BEFORE THE FEDERAL ELECTION COMMISSION

)
) MUR 6888
)

RESPONSE OF i360 TO THE COMPLAINTS IN MUR 6888

By and through the undersigned counsel, i360 hereby responds to the Complaint and Supplemental Complaint ("the Complaints") in the above-captioned Matter Under Review ("MUR"). Because the Complaints fail to allege a violation of the Federal Election Campaign Act of 1971 (as amended) ("the Act") or the Commission's regulations implementing same, and misstate both the law and the facts, we respectfully request that the matter be dismissed and the file closed:

- The Complaints have not alleged that i360 meets the payment prong of the coordination analysis; nor do they identify specific communications by i360 that would satisfy the content prong.

- i360 is not a common vendor: it did not create, produce, or distribute the communications vaguely referenced in the complaint; it did not use or convey any of its campaign customers' plans, projects, activities, or needs to its outside group clients; it employs both an organizational structure and a firewall to prevent such an occurrence; and its data does not contain communicative information, such as source, about its origin or use.

- i360 does not select the audience for its customers' communications; the customers themselves must decide on their content, timing, intended audience and the like; there is no information available to a customer regarding the activities of other customers.

- The Complaints make no allegations as to i360 regarding "Establish, Finance, Maintain, or Control" or in-kind contributions.

- The Commission has not previously concerned itself with data of this sort and instead has dismissed similar enforcement matters. Given that the Commission has previously approved a number of advisory opinions regarding list swaps, to change course now would improperly regulate a relatively new area by enforcement action rather than rulemaking.

I. INTRODUCTION

This matter stems from Complaints filed by the American Democracy Legal Fund, an IRS Section 527 political organization formed for the purpose of targeting Republicans with complaints. *See* ADLF IRS form 8871, Line 12. In the waning days of the 2014 election, ADLF's treasurer—who is also former communications director of the DNC and president of a Democrat-aligned Super PAC—issued press releases and filed the above-referenced Complaints against i360 and dozens of other respondents alleging a wandering grand conspiracy based on inference, innuendo, and conjecture that, as explained below, is unencumbered by either law or the facts.

Respondent i360 is a for-profit company that serves as a data warehouse and data resource vendor to its customers, which include businesses, not-for-profit entities, political committees, candidates, and political party committees. Palmer Aff. at ¶1 (attached hereto as Exhibit A). i360, at its core, is a commercial data vendor selling data from its data library of over 190 million voters and 250 million American consumers, including hundreds of aggregate data points on such individuals, as well as proprietary predictive modeling data. *Id.* at ¶2. In addition, i360 has developed grassroots and analytical tools that assist clients in managing and using data effectively, including data management platforms and mobile canvassing applications (so-called "apps").[1] *Id.* at ¶3.

---

[1] On a limited basis, i360 also provides other services, including media buying, but did not provide any such other services to any of the respondents in this matter, and those other services are not raised in the complaint and are beyond the scope of this current matter.

The Complaints center on a methodology the media has termed "Big Data" but which is more accurately described as "data analytics" and its product: "modeling." Over the past decade or so, and after the Commission's coordination rule was promulgated, data analytics have increasingly become a part of daily life—from internet searching to consumer habits to trends in the stock market—and political life is no exception. David W. Nickerson and Todd Rogers, "Political Campaigns and Big Data," Harvard Kennedy School at 4 (Feb. 2014), *available at* https://research.hks.harvard.edu/publications/getFile.aspx?Id=1040 (attached hereto as Exhibit B) [hereinafter "HKS Working Paper"].[2] While political campaigns and advocacy groups have always sought lists and information about voters and activists, data analytics and modeling on the scale we see today were virtually absent just ten years ago. Since then, data analytics have become big business and a regular part of campaigns across the political spectrum. *Id.*

The term "Big Data" is, at its core, misleading in this context. The essence of the modern data analytics is not just larger and better "lists" of information; modern techniques involve databases rather than lists. The difference is more than one of degree. Where the interchange of lists may involve exchanging information so that the recipient obtains new information it did not originally have (such as names, addresses or responses to various direct mail or other solicitations), data analytics starts from a different base: the entities using today's data analytics generally have

[2] This recent Harvard Kennedy School Working Paper "describes the state of contemporary data analytics" in campaigns. The paper was co-authored by David Nickerson, Notre Dame professor and former "Director of Experiments" of the Obama campaign's data analytics department, and Todd Rogers, Harvard Kennedy School professor and co-founder of the Analyst Institute, which focuses on developing best practices for "progressive political communications," and authored a 2008 after-action report for Democrat-aligned data vendor, Catalist.

the vast majority of the information utilized in generating those analytics already in their databases. The business of data analytics is not so much about obtaining new information as it is the analysis, interpretation and production of new and different information not previously available. This is generally done scientifically or through mathematical predictions or calculations.

What is exchanged today acts not as a source of new information, but a verification or test of algorithms and methodologies in use by the analysts themselves. Unlike list exchanges of a few years ago, the information exchanged today is not used directly by the analysts or their vendors as a means of communicating messages; instead it is used to validate already-existing choices and unrelated determinations. This is different from the sort of information considered by the regulations to be material strategic information about communications.

In other words, data analysis today is mathematics-based, and the influence of any single bit of information is very minor. In older, list-based methodologies, a campaign, for example, might try to obtain a list of persons who responded to a particular communication, about which the campaign had no previous information. Today, however, the data analysts already have thousands of data points on each individual, no single one of which is determinative.

The impact of data analytics is in the aggregation of these huge numbers of data points, and the resultant predictions, based on algorithms rather than past behavior, gives a more precise picture of the persons to be reached. Thus, the data obtained through information exchanges between databases is used not for targeting communications or determining strategies (except in the grossest sense);

it is used to validate or challenge the algorithms and predictive "models" that make up the proprietary products of the data analysts. The algorithms themselves, and the "modeled" data provided to customers as a result of those calculations, do not depend on the results of individual exchanges such as those used in prior years' list-based exchanges or rentals.[3]

One thing that has not changed with the recent data analytics boom, though, is that "[t]he foundation of voter databases is the publicly available official voter files maintained by Secretaries of State." HKS Working Paper at 8. Such voter files can and are enhanced with additional publicly available data such as geographic attributes, precinct-level turnout, census and other additional demographic data, as well as commercially available consumer data (which can provide information about education status, home ownership, consumer tastes, and thousands of other data points). *Id.* at 9-10. Commercial services can also match or "append" telephone numbers and email addresses to individual records to further enhance the library of data. *Id.* Through this process, any one individual could be associated with thousands of data points, such as "female," "age 47," "registered voter," "Smith County resident," "holds college degree," "married," "home owner," "subscriber to

[3] At the risk of using analogy (but described in more detail below), i360 furnishes a constantly-replenished warehouse of possible ingredients to choose from, but it does not dictate what is for dinner, who is invited or at what time the meal is served. Or to choose a more technological example, a campaign and independent advocacy group may both use Google to search for public domain images of a flag to use in a mailer, and Google uses both users' resulting "clicks" from that search to refine its algorithms about which images to display in the future, but, since there is no exchange of information about the uses of the image, that analysis and use of the results of the analysis does not amount to coordination. i360's service can also be viewed in the same light as Lexis or WestLaw – each contain massive amounts of information, but it is the end-user who must decide what to pull, what to emphasize and how to use it in an end-product. Merely because multiple users employ such services does not mean they are working in concert.

Field & Stream magazine," "voted in the 2012 primary," etc. Though this information is available to the public, some of it is by no means free and on its own is probably of very little of use to anyone. *Id.*

From such publicly available files, businesses like i360 have built proprietary predictive models, the backbone of which is information about a person's identity, geographical location, demographics, voting history, and other behaviors. *Id.* This raw data can be analyzed, modeled, and tested, so that groups such as i360 can create predictive models based on this data, where future behaviors and preferences can be anticipated by i360's proprietary modeling algorithms. *Id.*

Predictive models use statistical principles to test data and attempt to discern whether certain variables (the data points) suggest a predisposition to certain behaviors.[4] Commercial businesses such as i360 collect and purchase, organize, maintain, update, and aggregate individual data points into a commoditized database and develop predictive statistical models based upon that data using advanced data mining techniques and statistical pattern recognition.[5]

[4] For example, a predictive model could suggest that male homeowners over age 40 in a rural area who subscribe to *Field & Stream* are more strongly predisposed to purchase fishing equipment in the next six months than single male recent college graduates who live in an urban area and subscribe to *Hipster*. In the election context, variables, coupled with data such as previous turnout, exit polling, and other factors can be tested and analyzed to create predictive models concerning likely strength of association with a political viewpoint, propensity to vote, or other factors. The same is true in the fields of issue and legislative advocacy, where a person's buying habits and other factors help predict predisposition to certain messages.

[5] The HKS Working Paper authors, familiar with Obama for America's massive data analytics operation, explain this in terms of an individual campaign purchasing and aggregating this information—an operation simply unfathomable for the overwhelming majority of non-presidential campaigns to undertake in-house. Data vendors—both those catering to either side of the aisle, across the political and ideological spectrum, and those offering services to all who pay—fill this void for campaigns, organizations, and other smaller-scale customers.

This is no small matter; i360 employs a number of data scientists to maintain the database and proprietary modeling algorithms based on a multitude of individual data points. As in any statistically-based endeavor, there is a substantial potential for error or distortion, and thus a powerful corollary incentive for obtaining as much data as possible, and then repeatedly testing and verifying the algorithms and their products. Generally, in the study and business of statistical analysis and predictive modeling, the more valid data points are available to be tested and manipulated, the better.

Because it has built a very large and very sophisticated database and set of modeling and analytical tools, i360 has a variety of for-profit, non-profit and political customers who have varying needs for various data and data management. Some customers access the i360 library of enhanced data and predictive modeling through a web-based data management system. This system provides customers an interface to query the database based on attributes the customer desires to generate information.[6]

Critically, however, the database and its products are not determinants of communications strategy or usage; they are tools for use once a communications strategy is already determined. A customer must first decide on a preferred message, and also decide what sort of audience ought to receive their message. Once a customer has made such decisions, they can then query the i360 database. These queries are based, in large part, on the proprietary i360 predictive models,

[6] For example, a customer could pull or select a list of all female registered voters in Baltimore. A customer could further refine their pull to include all female registered voters in Baltimore who are registered Democrat yet appear to be pro-life under the age of 35.

although some customers can adapt their queries based upon customer-selected criteria. In other words, it is the end-user customer who ultimately decides what to select from the i360 data library, and usually does so on the basis of predictive models, and not the results of information exchanged as erroneously alleged in the Complaints. **Critically, in no case can a customer select data that that customer knows has been generated by a particular customer.** Palmer Aff. at ¶ 6. That information is simply not available to other customers, except as a result of the intensive processing of data as part of the modeled scores.

i360 also provides customers with data management tools. Several of i360's campaign customers (who are also respondents in the current matter) merely use this capability. This is known as "siloed" data; that is, data is kept in segregated "tables" or databases accessible only by that particular customer and its account representative. *Id.* at ¶ 5. Those customers upload their own data, where it remains in their own customer-specific siloed database. Such customers can then access and manipulate that data, and select from it based upon their needs and desires. For this, they generally pay a monthly fee and consent to allowing i360 to analyze the data. In other words, i360 is permitted to use this customer data to further refine and tag i360's already-existing data library and the predictive models it contains. i360 data analysts (who have no direct contact with campaign or non-campaign customers) perform this refining process by "testing" these imported data points, *i.e.*, comparing them with the results of its predictive models in i360's existing database. Though much of this imported data is redundant with existing data points, it is still valuable for i360's future purposes, because even redundant data enhances the i360 database and models with issue and demographic information relevant to individuals i360 already has in its database. In short, the new data is used principally to check (or "verify") if it fits the constantly evolving predictive models; if it doesn't, the models are reviewed and altered to account for any discrepancy or shifts over time.

Consistent with what scholars have observed, "[t]he vast majority of these variables literally do nothing to increase the power of the models of mass behavior once prior behavior is accounted for (*i.e.*, any power of income or education measures to predict voter turnout are subsumed by controlling for prior turnout)." HKS Working Paper at 10. More importantly, "the attributes of these individuals can be used to develop behavior scores to identify others who may be likely [to do the same]." *Id.* In other words, data uploaded by a customer into their own siloed database is not then passed on to other customers in its raw form. Instead, it is used by i360 to double-check and enhance its already existing data and modeling, and predict the behavior of others similarly situated. **The verification process does not transmit communicative strategy or information; it is used to refine and enhance the data produced by the predictive models.**

The agreement between i360 and GOP Data Trust LLC ("Data Trust") serves a similar principal function—testing and verifying existing data models of i360's already-robust and thorough pool of data. The Complaints do not deal with this element of data analytics—which as shown above, is the defining aspect of the behaviors they object to—because they fail to understand it or how the law applies to it. Instead the Complaints try to fit superficial media stories into a pre-ordained, obsolete narrative of a business model that does not apply here or elsewhere in the large data analytics world.

As it does with other commercial data resources, i360 entered into a business deal with Data Trust, another commercial data vendor[7] which possessed commoditized data that could be useful to the existing i360 data library and predictive models, and thus improve i360's product. Palmer Aff. at ¶ 7. Like all other data points in the i360 library and modeling database, end users cannot glean any identifying information about the source of any given data point. *Id.*. Nor can they ascertain who else may have used the data, how many times, for what purpose, or the like. *Id.* at ¶¶ 6-7.

The Complaints, by contrast, paint an inaccurate factual picture, and attempt to concoct a heretofore-unknown and erroneous theory of coordination. Under the misguided fictions contained in the Complaints, campaigns upload their own targeting lists and suggestive information to i360, which can instantly be accessed by otherwise-independent advocacy groups. Similarly, the Complaints speculate that the Republican National Committee moves information through Data Trust, to i360, which in turn makes it instantly available to others. This is flat wrong.

When a customer of i360 uploads information, it remains siloed in a customer-specific database, which other customers cannot access. The account representative who handles that particular account is firewalled from speaking with

[7] i360 entered into a business deal with Data Trust as an independent, for profit business enterprise. i360 lacks any first-hand knowledge regarding the establishment of Data Trust, or any of the accusations raised in the Complaints regarding Data Trust's relationship with the Republican National Committee.

others in i360 who handle other non-campaign customers. Palmer Aff. at ¶ 5. The only information updates available in real-time to customers are thosein their own siloes that such customers have themselves inputted *via* the data management system. Thus, when another customer uses the i360 data warehouse, it does so without any indication where its selected data came from: publicly available voter file, consumer data, other publicly available data, or data points generated by other customers. Critically, there is no communication between customers of material strategic information in this use, and, contrary to the unfounded conspiracy theory espoused in the Complaints, the use of such data is done at the sole discretion of the customer's strategists and communications personnel.

In sum, the Complaints' presentation of the facts and accompanying coordination conspiracy theory is wrong. The vast majority of what is used by independent advocacy groups comes from publicly available information, and has gone through several filters, performed by professional individuals who are not involved in any way in a customer's development of their plans, projects, activities or strategic needs, with the added protection of an anti-coordination firewall. What is left is mathematically reliable data, without any of the sort of human intuition or communication, strategic significance or any of the sorts of plans or needs contemplated by the Commission's coordination regulation. Simply put, there is thus no material strategic information about communications in the exchange of data in today's database-oriented world. The volume of information exchanged is not material (especially in the context of the overall information already in the database), the information is not used for strategic purposes (that is the focal point of the predictive models, not the data itself), and the information exchanged is not available for the purposes alleged in the Complaints. Ultimately, what is left in the i360 database is akin to an enhanced phone book, which does not instruct the end-user on who to contact or how to use it. Such decisions are left to the end-user customer.

The Complaints simply do not understand this aspect of the activities they complain about, and they therefore misstate the law as much as they misstate the facts. Using a list-based analysis to challenge a database-oriented system overstates some aspects and underplays others, and the combination is far from accurate, complete or illuminating. It is simply a recitation of opinions without any recognition of what is really going on.

II. ANALYSIS

A. The Complaints Misstate and Misapply the Coordination Rule

The Commission's coordination regulation, found at 11 C.F.R. § 109.21, sets forth a three-part test, commonly summarized as (1) the payment prong, (2) the content prong, and (3) the conduct prong. All prongs must be satisfied in order for the regulation to apply. *See* Explanation & Justification for Regulations on Coordinated and Independent Expenditures, 68 Fed. Reg. 421, 426 (Jan. 3, 2003) [hereinafter "2003 E&J"] ("For a communication to be 'coordinated,' all three prongs of the test must be satisfied" and "no one of these elements standing alone fully answers the question.").

This is true even at the initial reason to believe stage. *See* MUR 5823 (Club for Growth, Inc.) Factual & Legal Analysis at 8, 9 (finding no reason to believe despite the fact that advertisements "appear to meet both the payment prong and the content standard" because they "do not appear to meet the common vendor conduct standard."). The Commission has made clear that "reason to believe" is a heightened standard:

> The Commission may find "reason to believe" only if a complaint sets forth sufficient separate facts, which, if proven true, would constitute a violation of the FECA. Complaints not based upon personal knowledge must identify a source of information that reasonably gives rise to a belief in the truth of the allegations presented.

MUR 4960 (Hillary Rodham Clinton for U.S. Senate Exploratory Committee, Inc.), Statement of Reasons of Commissioners David M. Mason, Karl J. Sandstrom, Bradley A. Smith and Scott E. Thomas at 1 [hereinafter "MUR 4960 (Clinton) Statement of Reasons"]. *See also* MURs 5878 (Arizona State Democratic Central Committee), Statement of Reasons of Commissioners Donald F. McGahn, Caroline C. Hunter and Matthew S. Petersen at 4-7 (discussing Commission's treatment of heightened reason to believe standard); 4850 (Committee to Re-Elect Vito Fossella / Deloitte & Touche, LLP), Statement of Reasons of Commissioners Darryl R. Wold, David M. Mason and Scott E. Thomas at 2 ("A mere conclusory allegation without any supporting evidence does not shift the burden of proof to the respondents."). The Commission has consistently dismissed complaints claiming coordination that are sparse on facts or heavy on circumstantial evidence. *See* MURs 6611 (Friends of Laura Ruderman); 6368 (Friends of Roy Blunt); 6570 (Berman for Congress); 6359 (Voters Response); 6038 (Lamborn); 6077 (Coleman); 6050 (Boswell for Congress); 6059 (Sean Parnell for Congress); 6056 (Protect Colorado Jobs, Inc.); 5845 (Citizens for Truth); 6164 (Sodrel); 5754 (MoveOn.org Voter Fund); 5568 (Empower Illinois); 5576 (New Democrat Network); 5609 (Club for Growth, Inc.); and 5691 (Whalen).

1. Neither the "Payment" nor "Content" prongs are met

The Complaints fail to point to any communications made by i360 that could meet the payment prong. i360 does not pay for any communications and did not otherwise distribute communications at issue in the Complaints, and therefore does not satisfy the payment prong of the coordination analysis. 11 C.F.R. § 109.21(a)(1). Thus, the Complaints ought to be dismissed as to i360 for this reason alone.

Further, the Complaints fail to point to any particular communications they allege are coordinated. Rather, the Complaints attach what appears to be a random sample of undifferentiated communications. The coordination regulations, though, require more. *See* 11 C.F.R. § 109.21(a) (defining the scope of the regulation to a specific communication, *i.e.*, "A communication is coordinated . . ."); MUR 6077 (Coleman) First General Counsel's Report at 8 (requiring the complaint to allege "specific communications(s) . . . have been coordinated") (emphasis added); MUR 5845 (Citizens for Truth), Factual & Legal Analysis at 3 ("The complaint provides no information indicating whether the content prong may be satisfied."). The regulations center on the communication itself, requiring each prong of the analysis to be satisfied. 11 C.F.R. § 109.21(a). Because the regulations are communication-specific, a generalized accusation of atmospheric "coordination" is not enough under the coordination regulations or under the reason to believe standard. *See* MUR 4960 (Clinton) Statement of Reasons at 3 ("[P]urely speculative charges, especially when accompanied by a direct refutation, do not form an adequate basis to find reason to believe that a violation of the FECA has occurred"). Moreover, it appears that, to the extent any particular communications can be gleaned by inference from the Complaints, they are television advertisements. Since i360 is in the business of selling enhanced data with respect to the named Respondents—akin to an enhanced phone book based upon publicly available voter rolls and consumer data—the Complaints' accusations regarding that data have nothing to do with the television ads referenced in the Complaints. The Complaints can also be dismissed on this ground alone.

2. The "Conduct" prong is not met

The Complaints hinge on the erroneous notion that i360 serves as a "common vendor" under 11 C.F.R. § 109.21(d)(4). That section states:

> All of the following statements in paragraphs (d)(4)(1) through (d)(4)(iii) of this section are true:
>
> (i) The person paying for the communication, or an agent of such person, contracts with or employs a commercial vendor, as defined in 11 CFR 116.1(c), to create, produce, or distribute the communication;
>
> (ii) That commercial vendor, including any owner, officer, or employee of the commercial vendor, has provided any of the following services to the candidate who is clearly identified in the communication, or the candidate's authorized committee, the candidate's opponent, the opponent's authorized committee, or a political party committee, during the previous 120 days:
>
>> (A) Development of media strategy, including the selection or purchasing of advertising slots; (B) Selection of audiences; (C) Polling; (D) Fundraising; (E) Developing the content of a public communication; (F) Producing a public communication; (G) Identifying voters or developing voter lists, mailing lists, or donor lists; (H) Selecting personnel, contractors, or subcontractors; or (I) Consulting or otherwise providing political or media advice; and
>
> (iii) This paragraph, (d)(4)(iii), is not satisfied if the information material to the creation, production, or distribution of the communication used or conveyed by the commercial vendor was obtained from a publicly available

source. That commercial vendor uses or conveys to the person paying for the communication:

> (A) Information about the campaign plans, projects, activities, or needs of the clearly identified candidate, the candidate's opponent, or a political party committee, and that information is material to the creation, production, or distribution of the communication; or
>
> (B) Information used previously by the commercial vendor in providing services to the candidate who is clearly identified in the communication, or the candidate's authorized committee, the candidate's opponent, the opponent's authorized committee, or a political party committee, and that information is material to the creation, production, or distribution of the Communication.

The Commission was clear in 2003 when it instituted this regulation that "[t]he common vendor rule is carefully tailored to ensure that all . . . conditions be met." 2003 E&J at 436; *see also* MUR 6277 (Kirkland), Statement of Reasons of Commissioners Caroline C. Hunter, Donald F. McGahn and Matthew S. Petersen at 8 ("the common vendor standard cannot be met where there is no common vendor"). But with respect to i360, the conditions are not met, specifically because:

- i360 was not hired to create, produce, or distribute communications, and thus cannot use any other client's information in the creation, production or distribution of a communication.
- The data in the i360 data library or predictive models does not and cannot contain communicative information or any indicia of another customer's plans, projects, activities or needs.
- Customers' data is initially siloed so that no one customer may access another's data.
- i360 account managers (who are aware of a specific customer's data) for outside groups are firewalled from those for candidate or party groups.

a. i360 is not a "common vendor" because it was not hired to create, produce, or distribute any communications

Contrary to the Complaints' speculative and unfounded conclusory allegations, i360 does not come within the regulatory definition of "common vendor." Simply put, i360 was not hired by any of the other Respondents to either create, produce, or distribute any communications. Palmer Aff. at ¶ 8. Instead, i360 is in the business of data and the management thereof, and with respect to the respondents in this matter, does not "distribute" communications under the regulation. The Commission has already confirmed in its 2003 regulatory Explanation and Justification that the coordination regulation does not apply to such commercial vendors such as i360. There, the Commission made clear that the "standard only applies to a vendor whose usual and normal business includes the creation, production, or distribution of communications, and does not apply to the activities of persons who do not create, produce, or distribute communications as a commercial venture." 2003 E&J at 436; *see also FEC v. Christian Coalition*, 52 F. Supp. 2d 45 (D.D.C. 1999) (speaking to the meaning of "distribute" in the coordination context); Supplemental Notice of Proposed Rulemaking on General Public Political Communications Coordinated With Candidates, 64 Fed. Reg. 68951, 68951 (Dec. 9, 1999) (the proposed rules were intended to "incorporate . . . the standard articulated . . . in the *Christian Coalition* decision"). Data is not a communication, especially when it is anonymized as it is by i360 and most other modern data analytics organizations.

Past enforcement actions make the same point. For example, in MUR 6077 (Coleman), the Office of General Counsel observed that "a vendor is a 'common vendor' for the purposes of the Act only if the same vendor creates or distributes the ad alleged to be coordinated." MUR 6077 (Coleman), First General Counsel Report at 7. Here, there is not even an allegation that i360 either created or distributed any of the communications referenced in the Complaints. The same was true in MUR 6050 (Boswell for Congress). There, even where a campaign and an advocacy group used a common mail house, which in turn used the same postal indicia on each client's mailing, coordination allegations were dismissed. *See* MUR 6050 (Boswell for Congress), Factual & Legal Analysis at 9-10. Thus, because i360 did not distribute any communications, the Complaints ought to be dismissed.

b. Prior enforcement matters confirm that i360's business is not the sort from which common vendor coordination results.

Other Commission enforcement matters confirm that i360's business does not raise coordination concerns. For example, in MURs 5774 and 6038 (both of which concerned Lamborn for Congress), the Commission considered whether it was permissible for a commercial vendor to provide the same enhanced voter list already provided to the Lamborn campaign to an otherwise independent advocacy group. Further complicating the Lamborn matter was the fact the vendor was owned by the campaign manager, who in turn directed his company to provide to the advocacy organization the same enhanced list that was being used by the campaign. The Commission did not take issue with the arrangement, the result being that neither the vendor, the campaign, nor the advocacy organization violated the law.

16044391569

MURs 5774 and 6038 also confirm that any enhancement of data i360 receives from its political customers is of no legal consequence. It was made clear by the General Counsel in his initial report that the list at issue in MURs 5774 and 6038 was not simply an unaltered voter list, but instead had been enhanced by the campaign manager's company. Even with those enhancements—added under the direction and with the full knowledge of the campaign manager—coordination did not occur when the campaign-enhanced list was sent to an advocacy group. And as was the case in those MURs, here the bedrock of the i360 data library comes from publicly available sources, such as voter lists and consumer data. Thus, as was alluded to in MURs 5774 and 6038, this places it within the "public information" safe harbor found in the Commission's regulations. *See* 11 C.F.R. § 109.21(d)(2), (3) & (4)(iii) (regulation does not apply if "information material to the creation, production, or distribution of the communication was obtained from a publicly available source").

The same result ought to follow here. MURs 5774 and 6038 have already blessed analogous list sharing, even when a campaign manager specifically requests the list be given to an advocacy organization. That none of those more complicated facts are present here makes i360 a much easier case than that presented in MURs 5774 and 6038.[8] After all, i360 is not an old-fashioned list broker, nor does it

---

[8] In fact, some states not known for shyness in their efforts to regulate campaign finance have similarly recognized that data analytics is a compliant use of data and campaign finance information. *See, e.g.*, Minnesota Campaign Finance and Public Disclosure Board Advisory Opinion 418, at 3 (Oct. 4, 2011) ("Therefore, an association, including a for-profit association, may filter, improve or implement the use of Board data for a fee."), *available at* http://www.cfboard.state.mn.us/ao/AO418.pdf.

conduct the sort of "list development" contemplated by the Commission's regulations. *See* 11 C.F.R. § 109.21(d)(4)(ii)(G). The Complaints ought to be dismissed.

c. i360 is not a "common vendor" since the Complaints fail to allege any facts that i360 shared any sort of needs, plans or strategies among its customers

The Complaints incorrectly assume that merely by declaring "common vendor," they have sufficiently alleged impermissible coordination. Not so. As the Commission has made clear time and time again, mere use of the same vendor is not disallowed; vendors can provide services including those enumerated in 11 C.F.R. § 109.21(d)(4)(ii) to multiple customers without running afoul of the coordination rules:

> But under this final rule, even those vendors who provide one or more of the specified services are not in any way prohibited from providing services to both candidates or political party committees and third-party spenders. This regulation focuses on the sharing of information about plans, projects, activities, or needs of a candidate or political party through a common vendor to the spender who pays for a communication that could not then be considered to be made "totally independent" from the candidate or political party committee.

2003 E&J at 436-7. i360 does not share such information, and the Commission has already made clear that as a commercial vendor it can permissibly provide services to both candidate and political party clients and third-party spenders without triggering coordination. In fact, the regulations require much, much more: first, that the vendor also obtain and convey to others "[i]nformation about the campaign plans, projects, activities, or needs" of either the candidate or a political party; and second, that such information is used, and "material to the creation, production, or distribution of the communication." 11 C.F.R. § 109.21(d)(4).

The Complaints fail to offer any evidence or coherent analysis of such requirements. Instead, they offer a series of straw men, that they quickly torch. Relying on the resulting smoke, and in lieu of such requisite factual assertions, the Complaints weave a speculative conspiracy theory: that either the Republican National Committee or campaign committees are providing specific targeted lists that are then in turn provided instantly to otherwise independent advocacy groups *via* either i360, Data Trust, or a combination of the two. This is flat wrong on the facts. Simply put, i360 does not engage in substantial discussion with or otherwise obtain from its campaign customers information about plans, projects, activities or needs, and certainly does not then convey such non-existent information to others. Instead, i360 merely provides, houses and, in most instances, enhances data. In fact, the enhancement is so total in most cases that the exchanged data itself is not identifiable to anyone either directly or because it is such a small part of the modeled score that the customer receives.

Ultimately, the Complaints' conspiracy theory is easily debunked for two central reasons. First, any notion that information is being passed from a campaign or party committee to an independent advocacy group is false. Data of the sort at issue does not include such communicative information. i360 does not identify to its customers from where the data came, who else is using it, or for what purpose, or the like. Contrary to the Complaints' speculation, end-users do not—and cannot—know who discovered a particular piece of information, let alone when or how.

Second, the Complaints improperly attempt to lump in with campaign decision making and strategic communications, the concept of an enhanced data library (which finds its origin in publicly available information and qualitative statistical analysis). The existing coordination rules do not permit Complainants' creative leap. Rather, the common vendor and coordination regulations concern themselves with plans, projects, activities, needs—all things that are the subjective domain of the client, and which do not enter into the i360 library. i360's database constitutes a library of factual, qualitative information, available to be accessed and used by its clients on their own prerogative. Absent from this data is the sort of consulting, strategizing, or private campaign information from which coordination arises under FECA, the Commission's regulations, or the Commission's precedents.

d. In an abundance of caution, i360 maintains an internal firewall to avoid even the appearance of improper coordination

Even though i360 does not fall under the category of common vendor, nor does it obtain or convey the sort of insider information covered by the coordination rule, it nonetheless maintains an internal firewall, where those employees who deal with campaign or party committee customers do not deal with advocacy groups, and those who deal with advocacy groups do not deal with campaigns or party committees. *See* Exhibit C. Thus, such prudent business practices insulate i360 from any coordination allegation. *See* MUR 5823 (Club for Growth, Inc.), Factual & Legal Analysis at 6 ("Importantly, [respondents] assert that, as a matter of policy and practice, they isolate consultants or employees who also provide services to the candidates clearly identified in their advertisements (or their opponents and authorized committees)").

In addition, at the times relevant to this matter, i360 employs six account managers to handle its various customers. Those account managers who deal with campaign and party committee customers do not deal with relevant third-party advocacy customers, and those who deal with third-party advocacy customers do not deal with relevant campaigns and parties. Palmer Aff. at ¶5. They are physically separated, and sit in separate locations in i360's office. Under i360's policy, they are not permitted to speak to each other about their respective customer accounts. Thus, even if an account manager were to become privy to some potentially problematic insider information, the firewall policy prevents that information from going any farther. Any possible communicative link between a client's data and outputs of the i360 data library is broken not once, but twice, by the operation of both i360's firewalls and its business practices.

The data itself is not a means of transmitting private plans, projects, activities, and needs between campaigns and independent advocacy customers. Further, because any data points gleaned from the i360 data library or modeling scores have been scrubbed of source, circumstance, or any other identifying information, one customer cannot "reverse engineer" the data in the library in an attempt to gain information about another customer's activities, regardless of the Complaints' conjecture on the subject.[9] Because i360 does not engage in the sort of

---

[9] Although not relevant for the purposes of the MUR, i360 provides other services to other clients, such as media placement. To be clear, i360 did not provide such services to any Respondents in the current matter. Because of these other services that are not at issue here, i360's firewall is written in the broadest terms, and contemplates situations where i360 might be hired by others to place media. *See* i360 Firewall Policy (attached hereto as Exhibit C). Regardless of such details, however, the written firewall, distributed to i360 employees, creates a wall between data account managers to prevent precisely the sort of free flow of information imagined in the Complaints. The Commission

conduct covered by the Commission's coordination regulation, the Complaints ought to be dismissed.

B. <u>The Allegations Regarding "Establish, Finance, Maintain, or Control" and In-kind Contributions Between the Republican National Committee and the Data Trust Are of No Consequence to i360</u>

1. The Complaints do not claim that i360 was established, financed, maintained, or controlled by the Republican National Committee or any other political committee

The main thrust of the Complaints concern the relationship between the Republican National Committee, Data Trust and Crossroads. Other than what had been included in the Complaints, i360 has no knowledge or information regarding such accusations. In fact, any business relationship between i360 and Data Trust is still quite new, and occurred long after Data Trust was established.

Regardless, the Complaints ignored well-settled precedent on the issue, and create what can only be viewed as a novel approach to the issue, apparently based solely on a few board members supposedly having "ties to the Republican Party apparatus" (which has never been deemed a problem before) and an anonymous blog post, which by any measure does not establish reason to believe. *See* MUR 5338 (The Leadership Forum) Statement of Reasons of Commissioners Ellen L.

---

has never troubled itself with such technical details, and ought not do so here. *See* MUR 5823 (Club for Growth, Inc.), Factual & Legal Analysis at 12 (Respondents "fail[ed] to establish that they have a written firewall policy that was distributed to all relevant employees, consultants, and clients and, thus, do not meet the technical requirements for the safe harbor at 11 C.F.R. § 109.21(h). Nonetheless, their representations regarding the policy and practice of isolating common vendors sufficiently refute the speculative allegations of common vendor coordination, particularly when considered with the other information in their sworn statements." (footnote omitted)); *see also* MUR 5506 (EMILY's List) (proffer by counsel in initial response was sufficient to establish firewall and dismissal of matter even where group had field staff working directly with campaigns who were the subject of otherwise independent advocacy). Because i360 has established a firewall, the Complaints must be dismissed.

Weintraub and Scott E. Thomas at 2 (chiding OGC for selective use of newspaper articles); *see also id.* at 3 (criticizing "transitive theory of affiliation"); 52 U.S.C. §30109 (formerly 2 U.S.C. § 437g(a)(1)) (requiring that complaints be filed under oath).

2. Even if the Republican National Committee established, financed, maintained or controlled the Data Trust, that does not change the result with respect to i360

i360 does not read the Complaints as even attempting to allege sufficient facts to establish reason to believe as to the business relationship between the Republican National Committee and Data Trust. But even assuming *arguendo* that there are sufficient facts to question the relationship between the Republican National Committee and Data Trust, that does not change the result with respect to i360. For decades, the Commission has never taken issue with list exchanges, and never so much as hinted that such conduct could somehow raise either coordination or other issues. For example, since at least 1981, the Commission has recognized the commoditized nature of lists. *See* AO 1981-46 (Dellums) ("no contribution or expenditure would result and the transaction would not be reportable under the Act" where "a corporation exchanges names with [a] Committee"). In fact, the Commission has even made clear that national party committees can sell or lease (and thus exchange) their lists—even with 527 organizations, 501(c)(4)s, and labor organizations. *See* AO 2002-14 (Libertarian National Committee). Thus, even if the Republican National Committee and the Data Trust are one and the same (a point the Complaints fail to establish, but we only assume *arguendo*), that does not preclude or otherwise call into question the commercial business arrangement between i360 and Data Trust. It is already protected by a Commission advisory opinion. *See* 52 U.S.C. § 30108(c) (formerly 2 U.S.C. § 437f) ("Any advisory opinion rendered by the Commission under subsection (a) may be relied upon by . . . any person involved in any specific transaction or activity which is indistinguishable in all its material aspects from the transaction or activity with respect to which such advisory opinion is rendered."). In other words, even if there might be some speculative questions or other conjecture about the Republican National Committee and Data Trust, that does not preclude the dismissal of i360.

3. The Complaints' allegations concerning a possible in-kind contribution by the Data Trust to the Republican National Committee in the form of data services does not involve i360

The Complaints allege that the business deal between the Data Trust and the Republican National Committee may somehow have resulted in an excessive in-kind contribution to the Republican National Committee. The Complaints make no allegation or even suggestion that this claim involves i360 in any way. Nor could they. i360 engaged in a *bona fide* commercial business deal with the Data Trust. The Commission has a long history of approving *bona fide* business deals and industry practices, especially when it comes to voter, mailing, and other lists of individuals—even when it involved a federal committee—rather than viewing them as contributions or in-kinds. *See, e.g.*, AOs 1979-36 (Fauntroy) (approving industry standard practice regarding direct mail prospecting); 1981-46 (Dellums) (approving industry standard list exchange of names on mailing lists); 2002-14 (Libertarian National Committee) (approving party committee list rental per industry standards). Because the Commission has not previously questioned such commercial deals before, it ought not do so here, and the Complaints ought to be dismissed.

C. The Commission Ought to Decline the Invitation to Engage in "Regulation *via* MUR"

1. The FEC has not previously declared commercial transactions regarding data to be suspect

The Commission has considered a number of enforcement matters in which coordination is alleged. But a review of those matters confirms that the sort of commoditized information library at issue in the Complaints simply is not the sort of activity traditionally considered to raise questions regarding coordination. Nothing in the Commission's regulations, rulemakings or history suggests that this sort of commercial data activity creates coordination concerns—in fact, it has said precisely the opposite in a case that included allegations of request or suggestion by a campaign manager acting through a common vendor which he managed, *see* MURs 5774 and 6038 (Lamborn), and the Commission has already acknowledged the practice. *See* Notice of Proposed Rulemaking on Mailing Lists of Political Committees, 68 Fed. Reg. 52531, 52533 (Sept. 4, 2003) ("The Commission also seeks comment on whether it is usual and customary in the commercial list marketplace for one entity to provide raw list data to another entity that updates and enhances the data and where both entities consequently have access to the list."). The Lamborn MURs only confirmed the Commission's settled approach to the subject both in the regulation (first announced at its time of promulgation in its Explanation and Justification), and its history of approving advisory opinions, such as the approval of the Libertarian Party's proposal to rent its list to 501(c)(4) and 527 groups in AO 2002-14.

2. The advances in "Big Data" came about after the coordination rule was crafted.

The Commission ought not to be suspicious of Big Data. Campaigns and advocacy groups have long had methods of trying to predict the behavior of voters—from shoeboxes of index cards that served as a voter file to basic databases with scores assigned from 1 to 4. Beginning with the Democratic National Committee's utilization of the expertise of direct mail expert Matt Reese and the associated firm Claritas in the 1970's, predicting the behavior of voters, citizens, customers and consumers has evolved. Most recently, the use of Big Data as a means to target, analyze and predict voter behavior has become common across the political spectrum with Democratic companies and groups playing a leading role over the last few years in developing these techniques.[10] Merely because the advance of technology has changed and made more sophisticated the means available to target voters does not make these advances illegal.

It is likewise critical that the Commission recognize that similar advances in Big Data have been employed across commercial contexts. Organizations in various

---

[10] *See* Robert Knight, Catalist, the Left's Secret Electoral Weapon, Outguns GOP, *Washington Times* (Sept. 15, 2014), http://www.washingtontimes.com/news/2014/sep/19/knight-the-lefts-mighty-catalist/?page=all. In fact, the Complainant was created by David Brock of Media Matters, a group funded by George Soros, who also funds Catalist, the Democratic data company. *See* Kenneth P. Vogel, Media Matters' David Brock Expands Empire, *Politico* (Aug. 31, 2014), http://www.politico.com/story/2014/08/david-brock-citizens-for-responsibility-and-ethics-in-washington-110003.html (discussing Brock's "golden touch" with George Soros); George Soros Reinvests in Progressive-cause Data Company, *CNN* (Nov. 14, 2013), http://www.politico.com/story/2014/08/david-brock-citizens-for-responsibility-and-ethics-in-washington-110003.html ("A data analytics company specializing in progressive causes is getting a new round of investment funding, including $2.25 million from liberal billionaire George Soros, the company said Thursday.... Soros ... was an initial investor of Catalist's in 2006.").

industries gather, purchase, sell, enhance, amalgamate, exchange, share, and use data much like that at issue in this matter in an every-day business context. The Commission should not preclude those who wish to exercise their First Amendment rights or otherwise engage in politics from availing themselves of the same technologies and methods. To do otherwise would put those whose livelihood focuses on the sale of such data at a significant competitive disadvantage as compared to those with a broader focus. Instead, the Commission should ensure that commercial conduct that is standard practice for Amazon or Facebook or *The Washington Post* remains an available tool, technique, and technology for those who participate in elections and wish to contact citizens to encourage civic participation.

After all this, however, in all critical respects, Big Data changes nothing about the key aspects of coordination—even with this emerging technology, the speakers themselves still must devise strategies, plans and messages on their own. They must still determine what they want to say and do on their own, before they utilize Big Data to inform their own list of targeted individuals. Ultimately, the content of the message is determined by the end-user, as is the intended audience. All so-called Big Data provides is a more detailed and nuanced ability to find that intended audience.

3. The Administrative Procedure Act, Due Process and the First Amendment preclude the FEC from changing course here

The Act is clear that any new rule must be done *via* public rulemaking. 52 U.S.C. § 30108(b) (formerly 2 U.S.C. § 437f(b)) ("Any rule of law which is not stated in this Act or in chapter 95 or chapter 96 of title 26 may be initially proposed by the Commission only as a rule or regulation pursuant to procedures established in section 30111(d) of this title."); *see also* 5 U.S.C. § 551, *et seq.* (Administrative Procedure Act). Numerous Commissioners have repeatedly cautioned against using the enforcement process to make new rules or otherwise avoid so-called "regulation *via* MUR." *See* MURs 6206 (BASF Corporation), Statement of Reasons of Commissioners Matthew S. Petersen, Caroline C. Hunter and Donald F. McGahn at 2-3 (declining to find RTB because it "would have required us to rely on mere speculative inferences and to craft a new rule that goes beyond the plain language of the Act and Commission regulations"); 6113 (Kirby Hollingsworth), Statement of Reasons of Commissioners Matthew S. Petersen, Caroline C. Hunter and Donald F. McGahn at 9 ("As we have repeatedly explained, the enforcement process of the Commission is not the place to articulate new legal prohibitive norms"); 5937 (Romney for President, Inc.), Statement of Reasons of Commissioners Matthew S. Petersen, Caroline C. Hunter and Donald F. McGahn; 5835 (DCCC / Quest Global Research Group, Inc.), Statement of Reasons of Commissioners Matthew S. Petersen, Caroline C. Hunter and Donald F. McGahn at 9 ("As we have done in other matters, we decline the invitation to use the enforcement process to make new law, and we will not engage in so-called regulation via MUR".); and 5541 (The November Fund), Statement of Reasons of Commissioners Matthew S. Petersen, Caroline C. Hunter and Donald F. McGahn.

In addition to such administrative restraints, the Supreme Court has made clear that administrative agencies cannot simply change course and launch either new or previously rejected legal theories in the enforcement context. *See FCC v. Fox Television Stations, Inc.*, 132 S. Ct. 2307 (2012). To do so violates fundamental notions of due process. *Id.* Similarly, given that the investigation of speculative coordination accusations involves the government oversight of the sort of speech that is at the core of First Amendment protection, in a close call, the Commission is obligated to err on the side of liberty. As the Supreme Court recently noted, "validly conferred discretionary executive authority is properly exercised . . . to avoid serious constitutional doubt." *Arizona v. Inter-Tribal Counsel of Ariz., Inc.*, 133 S. Ct. 2247, 2259, 570 U.S. __ (2013). Thus, the Commission is obligated to honor the doctrine of constitutional avoidance, which has a particularly urgent application here, a matter that concerns protected politics. *See Edward J. DeBartolo Corp. v. Fla. Gulf Coast Bldg. & Constr. Trades Council*, 485 U.S. 568, 575 (1988); MUR 5541 (The November Fund), Statement of Reasons of Commissioners Matthew S. Petersen, Caroline C. Hunter and Donald F. McGahn.; *see also* MUR 5879 (Harry Mitchell for Congress), Statement of Reasons of Commissioners Caroline C. Hunter, Donald F. McGahn and Matthew S. Petersen (discussing the constitutional significance of coordinated and independent expenditures).

III. CONCLUSION

The i360 business i360 conducts is not the sort that is contemplated by the regulatory common vendor restriction; i360 does not create, produce, or distribute messages. And because it is not a campaign, party, or outside group making communications, it does not engage in activities that could be considered contributions to federal candidates. FECA and the regulations cannot be read broadly to apply to a commercial vendor—one which does not make communications and which has taken appropriate and responsible measures to safeguard against being a means of coordination between their clients—in an unfounded attempt by a Complainant to somehow wrongly bring such activities under the purview of the Commission and the coordination rules.

Ultimately, the end-user customers of i360 determine what messages they wish to disseminate to whatever segments of the population those customer choose, at whatever time and means they choose. The use of i360 data does not provide such information, nor could it. It is the i360 customer who decides what they want to communicate in a message—and how they want to communicate their message—and selects what people they want to target with that message. Only then does the customer query the i360 data library in order to build a list and obtain contact information for that universe. And no record of what universes the customer pulls for what reasons and when could possibly reach any other customer of i360. In other words, i360 does not select the target audience; each customer does without knowing what any other customer is doing or has done.

For the foregoing reasons, Respondent respectfully requests that the Commission find **NO REASON TO BELIEVE** that a violation occurred, that this matter be **DISMISSED** and that the Commission **CLOSE THE FILE.**

Respectfully submitted,

Donald F. McGahn II
Jones Day
51 Louisiana Ave., N.W.
Washington, DC 20001
(202) 879-3939

Counsel for i360

Exhibit

B

HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Political Campaigns and Big Data

Faculty Research Working Paper Series

David W. Nickerson
University of Notre Dame

Todd Rogers
Harvard Kennedy School

Revised, February 2014
RWP13-045

Visit the **HKS Faculty Research Working Paper Series** at:
http://web.hks.harvard.edu/publications

Political Campaigns and Big Data

David W. Nickerson and Todd Rogers

David W. Nickerson is Associate Professor of Political Science, University of Notre Dame, South Bend, Indiana. He served as the "Director of Experiments" in the Analytics Department in the 2012 re-election campaign of President Barack Obama. Todd Rogers is Assistant Professor of Public Policy, Harvard Kennedy School of Government, Cambridge, Massachusetts. He co-founded the Analyst Institute, which uses field experiments and behavioral science insights to develop best practices in progressive political communications. Their email addresses are dnickers@nd.edu and Todd_Rogers@hks.harvard.edu.

ABSTRACT (145 words):

Modern campaigns develop databases of detailed information about citizens to inform electoral strategy and to guide tactical efforts. Despite sensational reports about the value of individual consumer data, the most valuable information campaigns acquire comes from the behaviors and direct responses provided by citizens themselves. Campaign data analysts develop models using this information to produce individual-level predictions about citizens' likelihoods of performing certain political behaviors, of supporting candidates and issues, and of changing their support conditional on being targeted with specific campaign interventions. The use of these predictive scores has increased dramatically since 2004, and their use could yield sizable gains to campaigns that harness them. At the same time, their widespread use effectively creates a coordination game with incomplete information between allied organizations. As such, organizations would benefit from partitioning the electorate to not duplicate efforts, but legal and political constraints preclude that possibility.

The all-encompassing goal of political campaigns is to maximize the probability of victory. To that end, every facet of a campaign is evaluated by how many votes an activity will generate and at what cost. But in order to perform this cost-benefit analysis, campaigns need accurate predictions about the preferences of voters, their expected behaviors, and their responses to campaign outreach. For instance, efforts to increase voter turnout are counter-productive if the campaign mobilizes people who support the opponent. Over the past six years, campaigns have become increasingly reliant on analyzing large and detailed data sets to create the necessary predictions. While the adoption of these new analytic methods has not radically transformed how campaigns operate, the improved efficiency gives data savvy campaigns a competitive advantage in targeting. This has led the political parties to engage in an arms race to leverage ever growing volumes of data to create votes. This paper describes the utility and evolution of data in campaigns.

As recently as a decade or two ago, the techniques used by political campaigns to predict the tendencies of citizens appear extremely rudimentary by current standards. At that time, citizens' likely support was gauged primarily by their party affiliation and the "performance" of the precinct in which they lived (that is, what percentage of the precinct had voted for a given party in the recent past). Whether a person was likely to turn out and vote was often based on the past four general elections; for example, it was not uncommon to hear phrases like "2 of 4 voter" or "3 of 4 voter" used in campaign targeting plans. Past donors would be recontacted and asked for a flat amount of money (or perhaps asked for their highest previous contribution if that information was available) and prior volunteer captains would be recontacted, but intermittent volunteers were unlikely to appear on any lists. At this time, a "numbers driven campaign" implied that candidates and their advisors paid close attention to poll numbers and adjusted policies in response to surveys. A memorable example of this dynamic is the story of President Clinton's advisor Dick Morris fielding a poll to choose Jackson Hole, Wyoming as the vacation spot for the president (Kuhn 2007). Presidential campaigns targeted states based on historical notions of which states could see the vote swing either way, combined with the realities of the campaign budget.

In retrospect, the reliance of political campaigns on such rough—although often useful—heuristics is puzzling. Campaigns a decade ago already possessed considerable information on citizens' preferences based on what they had collected directly from volunteers, donors, and their own polling. Voter registration rolls were available from Secretaries of State. Detailed census information was available. Why did campaigns take so long to realize the value of information resources they already possessed?

Part of the answer is technological: adequate storage and computing power required large investments and were beyond the infrastructure of nearly all campaigns and state parties. Even if an entrepreneurial campaign made that investment, much of the available data would not have been as reliable as it is today. States were not required to keep electronic copies of which citizens voted in each past election until 2002 with the passage of the Help America Vote Act of 2002 (42 U.S.C. § 15483), so using that data would have been onerous in many regions.

But perhaps the biggest impediment to wider adoption of data-driven campaigning was simply that statistical thinking – and the human capital that produces it – had not yet taken root in the world of political consulting. Campaign consultants generate most of their business through social networks and are judged by win/loss records. Political candidates are typically trained in non-quantitative fields like law, education, and medicine, and are more focused on fundraising and voter outreach than the nitty-gritty of managing a campaign. There were certainly consultants specializing in campaign data analytics, and the development of predictive scores existed as a niche business, but most campaign decisions did not rely on these approaches. There were too few people with the skills required to make a noticeable impact on how campaigns operated, and too few decision-makers equipped to appreciate the effect that a fuller use of information could have. At that time, mail vendors were on the cutting edge of using consumer data for modeling purposes and at least a decade ahead of the political campaign learning curve (Malchow 2003).

These impediments to data-driven campaigning have changed in recent years. The costs of purchasing, storing, managing, and analyzing data have decreased exponentially. The supply of quantitatively oriented political operatives and campaign data analysts has increased as predictive analytics has gained footholds in other sectors of the economy like banking, consulting, marketing, and e-commerce. To reduce the need for individual campaigns to spend scarce funds purchasing citizen information from commercial vendors, the national parties have decided to construct, maintain, and regularly augment their own voter databases (McAuliffe and Ketten 2008, p. 280-287).

These conditions have provided fertile ground for analytically-minded consultants to apply statistical tools to campaign activities and campaign data. Contemporary political campaigns amass enormous databases on individual citizens and hire data analysts to create models predicting citizens' behaviors, dispositions, and responses to campaign contact. This data-driven campaigning gives candidates and their advisors powerful tools for plotting electoral strategy. A political campaign has limited financial resources. It can use this data-driven approach to shape decisions about who the campaign should target, with a sense of how much such contact will affect voter preferences, behaviors like fundraising, or turnout at the polls. This technology allows campaigns to target campaign outreach tactically at particular individuals and then also to aggregate these predictive estimates up to the jurisdiction-level to inform large-scale strategic decisions.

Given that campaigns view their analytic techniques as secret weapons to be kept out of the hands of opponents, the public discourse on campaign data has been largely speculative and somewhat hypothetical, ranging from hyping the performance of the tools (Scherer 2012) to alarmist concerns about the personal privacy of voters (Duhigg 2012). This paper describes the state of contemporary campaign data analytics. We begin by explaining why campaigns need data and the "predictive scores" that they seek to calculate. We then describe where that data comes from and the techniques used to analyze political data. We conclude by noting several challenges facing campaigns as data analytics become more widely used and increasingly accurate. The analytics revolution has not radically transformed campaigns in the manner that television did in the 1960s, but in a close political contest, data-driven campaigning can have enough effect to make the difference between winning and losing.

Why Do Campaigns Need Data?

Contemporary campaigns use data in a number of creative ways, but the primary purpose of political data has been – and will be for the foreseeable future –providing a list of citizens to contact. Campaigns need accurate contact information on citizens, volunteers, and donors. Campaigns would like to record which citizens engage in specific campaign-supporting actions like donating money, volunteering, attending rallies, signing petitions, or expressing support for candidates or issues in tracking polls. Indeed, the Federal Election Commission requires campaigns and coordinated committees to disclose the identity of all individuals who contribute more than $200 during the calendar year. These disclosure requirements mean that campaigns have a legal requirement – as well as financial incentive – to maintain good lists of donors.

Campaigns also use data to construct predictive models to make targeting campaign communications more efficient and to support broader campaign strategies. These predictive models result in three categories of "predictive scores" for each citizen in the voter database: behavior scores, support scores, and responsiveness scores.

Behavior scores use past behavior and demographic information to calculate explicit probabilities that citizens will engage in particular forms of political activity. The primary outcomes campaigns are concerned with include voter turnout and donations, but other outcomes such as volunteering and rally attendance are also of interest.

Support scores predict the political preferences of citizens. In the ideal world of campaign advisers, campaigns would contact all citizens and ask them about their candidate and issue preferences. However, in the real world of budget constraints, campaigns contact a subset of citizens and use their responses as data to develop models that predict the preferences of the rest of the citizens who are registered to vote. These support scores typically range from 0 – 100 and generally are interpreted to mean "if you sample 100 citizens with a score of X, X percent would prefer the candidate/issue". A support score of "0" means that no one in a sample of 100 citizens would support the candidate/issue, "100" means that everyone in the sample would support the candidate/issue, and "50" means that half of the sample would support the candidate/issue. Support scores only predict the preferences at the aggregate-level, not the individual-level. That is, people with support scores of 50 are not necessarily undecided or ambivalent about the candidate/issue and, in fact, may have strong preferences. But when citizens have support scores of 50, it means that it is difficult to predict their political preferences.

Responsiveness scores predict how citizens will respond to campaign outreach. While there are theoretical rationales as to who might be most responsive to blandishments to vote (Arceneaux and Nickerson 2009) and attempts at persuasion (Hillygus and Shields 2008), in general, predicting which individuals will be most and least responsive to particular direct communications in a given electoral context is difficult. Campaigns can use fully randomized field experiments to measure the response to a campaign tactic (Gerber and Green 2000, 2008; Nickerson and Rogers 2010; Arceneaux and Nickerson 2010; Nickerson 2005; Nickerson, Friedrichs, and King 2006; Bryan, Walton, Rogers and Dweck 2011; Gerber and Rogers 2009; Bailey, Hopkins and Rogers 2013; Rogers and Nickerson 2013). The results of these experiments can then be analyzed to detect and model heterogeneous treatment effects (i.e., predictive scores) that guide targeting decisions (Issenberg 2012a, 2012b, 2012c). Some of the results of these experiments can only be used to inform decisions in future elections (e.g., the results of most voter turnout experiments necessarily come after Election Day), but others can be conducted during the election cycle to improve efficiency in real time. For example, the lessons from experiments evaluating the efficacy of treatments aimed at increasing observable behaviors like donations and volunteering can be put to immediate use. Similarly, the persuasiveness of campaign communications can be gauged through randomized experiments that measure voter preferences through post-treatment polling of the treatment and control groups. The citizens found to be especially responsive to the campaign treatment in these pilot experiments – as reflected in the responsiveness score – can be targeted during a larger roll out of the campaign treatment. Conversely, citizens who are unresponsive, or are predicted to respond negatively, can be avoided by the campaign.

Campaigns are primarily concerned with the practical question of how accurately predictive scores forecast the behaviors, preferences, and responses of individual citizens, not with testing an academic theory. As a result, the variables included in the construction of these scores often have thin theoretical justifications. That said, a variable in a data set that is found to predict an outcome of interest but has no theoretical rationale for the relationship is more likely to prove to be spurious when validated against an "out-of-sample" dataset. Thus, successful predictive scores need not be based on theories or imply causal relationships. However, campaign data analysts must think critically and creatively about what variables sensibly relate to their outcomes of interest in order to generate predictive scores with the external validity required by campaigns.

Where Does Campaign Data Come From?

Procuring and maintaining large databases of citizens with up-to-date information from multiple sources may seem straightforward, but it is a nontrivial logistical hurdle and requires substantial financial commitment. After all, people frequently change residences and contact information (Nickerson 2006a). Campaigns also need to track their own behavior to limit awkward interactions with citizens who have been contacted multiple times previously.

In the recent past, campaigns struggled to manage and integrate the various sources of their data. The data collected by those working on digital communications rarely linked with the data collected by those working on field operations--meaning canvassing, phone calls, volunteer recruitment, and so on---or fundraising. One of the most heralded successes of the 2012 campaign to re-elect President Obama was the creation of *Narwhal*, a program that merged data collected from these digital, field, and financial sources into one database (Gallagher 2012; Madrigal 2012). As a result, the Obama re-election campaign began with a 10TB database (BigData-Startups 2013) that grew to be over 50TB by the end of the election (Burt 2013).

The foundation of voter databases is the publicly available official voter files maintained by Secretaries of State, which ensure that only eligible citizens actually cast ballots and that no citizen votes more than once.[1] The official voter file contains a wide range of information. In addition to personal information such as date of birth and gender,[2] which are often valuable in developing predictive scores, voter files also contain contact information such as address and phone. More directly relevant to campaigns, certain details about past electoral participation are also recorded on official voter files. *Who* citizens vote for is secret, but *whether* citizens vote is reflected in official voter files – as is the method used to vote: for example, in person on Election Day, or by use of absentee or another form of

---

[1] The exception to this rule is North Dakota, which does not have a voter registration system.. Eligible voters simply show up and prove their eligibility by showing a valid ID, utility bill, or having a neighbor vouch for their residency.
[2] In states that were subject to the Voting Rights Act, the self-identified race of the registrants is included on official voter files, though this may change in light of the Supreme Court's June 25, 2013, ruling in *Shelby County v. Holder* 570 US ___ (2013).

early voting. This information concerning past vote history tends to be the most important data in the development of turnout behavior scores, which is unsurprising given that the act of voting reveals the person to be a person with a high propensity to vote.

The geographic location of citizens' residences can also provide valuable information, because campaigns can merge relevant Census and precinct data with the information on citizens in the voter database. Census data—such as average household income, average level of education, average number of children per household, and ethnic distribution—is useful for the development of a host of predictive scores. Campaign data analysts also append the aggregated vote totals cast for each office and issue in past elections in each citizen's precinct to individual voter records in the voter database. Even being mindful of ecological fallacy—that is, inferring someone's individual characteristics based on their membership in a larger group or cluster—this aggregate-level information in fact tends to increase predictive score accuracy.

Campaign data analysts also can append two types of data from consumer databases. First, and most essentially, they seek updated phone numbers. Phone calls are a critical feature of campaigns. While a volunteer knocking on doors will make successful contact with 2 – 4 people/hour, a volunteer making phone calls can reach 10–15 people/hour (Nickerson 2006b; 2007a). Using an automated dialer, the total can be even higher. While most official voter files contain phone numbers, they are often out of date and coverage is incomplete. Election officials only request a phone number from voters registering for the first time, and so if someone continues voting in the same jurisdiction over time, it's not uncommon to find numbers that are 20 years out of date. Because current phone numbers are so important, campaigns find it worthwhile to purchase more accurate contact information available from consumer data firms.

Campaigns can also purchase a wide range of additional information from consumer data vendors relatively inexpensively, such as estimated years of education, home ownership status, and mortgage information. In contrast, information on magazine subscriptions, car purchases, and other consumer tastes are relatively expensive to purchase from vendors, and also tend to be available for very few individuals. Given this limited coverage, this data tends not to be useful in constructing predictive scores for the entire population—and so campaigns generally avoid or limit purchases of this kind of consumer data. The vast majority of these variables literally do nothing to increase the predictive power of models of mass behavior once prior behavior is accounted for (i.e., any power of income or education measures to predict voter turnout are subsumed by controlling for prior turnout).

While campaigns do purchase some information, the vast majority of the useful information campaigns collect about individuals is provided by the individuals themselves. For example, those who have donated and volunteered in the past are high-value prospects for fundraising and volunteer-recruitment in the future. Moreover, the attributes of these individuals can be used to develop behavior scores to identify others who may be likely to donate or volunteer. Similarly, information about individuals who answered the phone or door in the past can be used to develop behavior scores for others who may be likely to be contactable moving forward. Data collected from online activities can be of particular value as well, because such activities require a relatively low thresholds for citizens to take action. For the small set of citizens who provide an email address to the campaign to receive campaign emails[3], all of their activity concerning those emails—for example, sign up, opening emails, clicking links in emails, taking actions like signing petitions—can be tracked and used to predict levels of support for the candidate or focal issue, likelihood of taking action, and in many cases the policy areas of greatest interest (for example, imagine a voter who opens emails about taxes twice as often as any other topic). Thus, a state party or political organization can compile valuable information for developing predictive scores just by maintaining accurate records of its interactions with citizens over time.

[3] In 2012, the Obama campaign had email addresses for 20 million supporters (Haberman 2013) compared with 13 million in 2008 and the 3 million addresses collected by the 2004 Kerry campaign (Vargas 2008).

In short, many of the claims about the information that campaigns purchase about individuals is overblown; little of the information that is most useful to campaigns is purchased. Official voter files are public records, census and precinct-level information are also freely available, and individual citizens themselves volunteer a wealth of data that can be used to develop scores that predict all citizens' behaviors and preferences. In fact, predictive scores can often allow campaigns to estimate some citizen preferences and behaviors more accurately than direct reports from citizens themselves (Rogers and Aida 2013; Ansolabehere and Hersh 2012). People may not be actively misrepresenting their intentions, but the desire to project a positive image of the self may lead voters to over-estimate the degree to which they will participate in a given election. Again, the most important piece of information campaigns purchase tends to be phone numbers – and this is purchased with the intent of performing the old-fashioned task of calling citizens directly. Because the most useful information tends to be collected directly from citizens, one of the most valuable data acquisition activities campaigns engage in is exchanging their information with that of other allied political organizations (when legal) to increase the breadth and scope of data that will useful for the development of predictive scores.

An interesting result of the type of data that campaigns acquire directly from citizens is that campaigns are able to predict with greater accuracy which citizens will *support* their candidates and issues better than which citizens will *oppose* their candidates or issues. Information regarding citizens who donate, volunteer, and subscribe to email lists is available to campaigns and can be used to predict which other citizens will be similar. In contrast, citizens who do not perform such behaviors at all, or who perform similar behaviors for opposing campaigns, cannot be directly observed, so discriminating among the citizens who do not actively support a campaign is a much more challenging task. As a result the distribution of support scores typically have 2 – 3 times more voters with the highest scores (99 and 100) than the lowest (0 and 1). This imbalance does not imply that the opposition enjoys less passionate support or the data analysts failed in their predictive task; it is a natural result of being able to observe the activity of only one campaign's supporters in an electoral competition. Similarly, because the foundations of voter databases are official voter files from states, campaigns tend to have much more information on citizens who have voted and are registered than citizens who have never voted and are not registered. Predictive models can still be constructed to predict fruitful geographies or people to target for registration drives, but the data available is much sparser and the models necessarily more coarse. This likely exacerbates the inequality in campaign communication and outreach between those who are already politically engaged and those who are not, and between voters and non-voters (Rogers and Aida 2013).

How Do Campaigns Analyze Data to Develop Predictive Scores?

The predictive scores campaigns construct can be roughly divided into two types. The first predicts the behavior or attitudes of voters. These models do not make any causal claim about why these individuals vote or donate or support the candidate; they merely predict the focal trait. As such causation is not a major concern and the goal of the analyst is primarily to avoid over fitting the data. The second type of score predicts how voters will respond to campaign outreach. These responsiveness scores typically come from exploring heterogeneous reactions to campaign treatments in randomized field experiments. The causal effect of the campaign outreach is established by the experiment and these estimated effects are used as parameters for strategic decision making. However, the moderators predicting strongly positive or weakly positive (or even negative) responsiveness to the treatment are not causal. In other words, the data may have been generated by an experiment, but the enterprise of modeling responsiveness to the treatment remains a matter of finding observed differences across types of subjects that predict large or small treatment effects. Thus, even the search for moderators of the treatment effect in an experiment is essentially observational in nature.

Most of the analytic techniques employed by campaign data analysts are taught in standard undergraduate econometrics or statistics classes. Currently, the vast majority of the predictive scores used by campaigns are created by a campaign data analyst (or a team of them) using simple regression techniques: ordinary least squares for continuous outcomes; logistic regression for binary outcomes; and, rarely, tobit for truncated data like dollars donated or hours volunteered. The skills necessary for developing such models are widespread and the models can easily be customized to specific political environments. For instance, party registration is not predictive of candidate preference for older citizens in many Southern states, because the South was historically solidly Democratic and remained so at the state level well after the civil rights movement transformed the national political environment. Campaign data analysts modeling candidate support in these states need to be attuned to contextual facts like this, and can then accommodate them in regression analyses.

There are two major downsides to using regression techniques for constructing campaign models. First, the utility of techniques that uncover correlations is highly dependent on the talent of the particular campaign data analyst employing them. A capable campaign data analyst who is familiar with the properties of the variables available in voter databases can generate highly accurate predictive scores for citizens. However, a slightly less capable campaign data analyst might generate predictive scores that are only slightly better than the unsophisticated methods employed by earlier campaigns. As an example, consider the task of predicting a person's likelihood of voting in an election. Controlling for the whole set of turnout history available (often more than 50 elections) will typically predict around a third more variance in individual turnout than the old "of 4" rule of thumb (i.e., did the person vote in 0, 1, 2, 3, or 4 of the past elections). However, these variables all tap into a common latent propensity to vote and exhibit considerable collinearity. As a result, the coefficient for several of these variables will be negative and statistically significant. There is no theoretical rationale for why turnout in one election would decrease turnout in a future election, so observing negative coefficients would suggest that the analyst has over fitted the data and should pare back the number of variables used or model the propensity to turnout differently. They will also construct relevant variables (e.g., past turnout among people in the household) and insert theoretically informed interactions (e.g., ethnicity of the voter by ethnicity of the candidate) to improve model fit. The marginal gains from these new variables are rarely as large as the initial gains from using a wide range of past turnout decisions, but that is to be expected – the gains from good predictive models is incremental. Since the people running campaigns rarely have experience or expertise in data analytics, the competence of the campaign data analysts they employ cannot be taken for granted.

The second drawback is that unique regression models typically need to be constructed for different regions, issues, and candidates, so the "modeling by hand" approach to analysis offers few economies of scale. While individual campaign data analysts likely become more efficient with each successive model they develop, constructing models for multiple races around the country either requires a small army of campaign data analysts, or else settling for very general national models that are not adapted for local contexts.

Thus, campaign data analysts have been seeking more systematic methods for selecting a preferred regression. The commercial marketing industry often uses a form of "machine learning" (for example, k-means clustering or k-nearest neighbor classifiers, see Gan, Ma, and Wu 2007) and other to divide consumers into categorical types like "blue collar, grilling, SUV owner." However, these statistical methods to group similar individuals or households are less useful for campaign data analysts because strategic cost-benefit decisions in campaign planning are based on individual-specific probabilities for particular outcomes. For example, knowing that a set of citizens are similar in many dimensions does not assist with targeting if those dimensions are not highly correlated with behaviors like voting, ideology, and propensity to donate. For this reason, supervised learning algorithms are typically more appropriate for the task of modeling political data.

Supervised machine learning includes methods such as classification and regression trees (Breiman et al. 1984). In a regression tree approach, the algorithm grows a "forest" by drawing a series of samples from existing data; it divides the sample based on where the parameters best discriminate on the outcome of interest; it then looks at how regressions based on those divisions would predict the rest of the sample and iterates to a preferred fit. The researcher chooses the number of "trees"—that is, how many times the data will be divided. In the particularly popular "random forests" algorithm for implementing a regression tree (Breiman 2001), the algorithm uses only a randomly drawn sub-set of variables in each tree to decide on the fit rather than the entire set of available variables. The payoff for this approach is that it generates estimates of what parameters are most important: that is, what parameters add the most predictive power when the group of other parameters is unchanged. Aside from its analytical advantages "random trees" is a popular decision tree ensemble algorithm because it has very few tuning parameters and is available as an R-package, so that analysts with little formal education in statistics can develop the models. Bayesian Additive Regression Trees have similar advantages (Chipman, George, and McCollough 2010; Green and Kern 2011).

Supervised machine learning presents three major advantages for campaign data analytics. First, these classes of estimators are typically non-linear, so commonly known nonlinear relationships—such as the curvilinear relationship between age and turnout (i.e., older cohorts vote at higher rates than younger cohorts but this relationship peaks among group 60 – 70 years old and then reverses) — are easily accommodated by the algorithms. Second, the approach involves less discretion for the individual campaign data analyst, so the quality of the predictive scores generated is not as heavily dependent on the capabilities and integrity of analysts. People constructing the models still need to input the most diagnostic variables and set up rigorous out-of-sample tests to validate the models, but the algorithms are written in advance and run identically for every citizen in the voter database. Finally, these data-mining algorithms are relatively scalable. Some techniques may be computationally intensive and the variables included may need to be customized, but generally the marginal cost of constructing additional models is lower using these algorithms than having a campaign data analyst construct new models from similar databases by building a series of regressions from the ground up.

The major downside of these regression tree algorithms from the campaign's perspective is that their use is relatively new and not widespread, and it will take experience to see how to trim the regression trees and customize the tuning parameters in a way that satisfies political requirements. Campaign data analysts must also take great care to not overfit their models to their data (Dietterich 1995), in which case the results become less likely to apply outside the model. Typically, there is not sufficient data from any single jurisdiction to create a unique model, so the data from several jurisdictions need to be pooled to produce useful predictive scores. Most algorithms can be adapted to accommodate jurisdiction-specific political requirements, but only a small fraction of campaign data analysts today have the necessary skillset. In sum, as campaign data analytics becomes more common, sophisticated, and mature, the techniques most widely used will likely move away from creating a judgment-based series of regressions to those based on customized machine learning algorithms like regression trees.

How Are Predictive Scores Used?

Campaigns use predictive scores to increase the efficiency of efforts to communicate with citizens. For example, professional fundraising phone banks typically charge $4 per completed call (often defined as reaching someone and getting through the entire script), regardless of how much is donated in the end. Suppose a campaign does not use predictive scores and finds that upon completion of the call 60 percent give nothing, 20 percent give $10, 10 percent give $20, and 10 percent give $60. This works out to an average of $10 per completed call. Now assuming the campaign sampled a diverse pool of citizens for a wave of initial calls. It can then look through the voter database that includes all citizens it solicited for donations and all the donations it actually generated, along with other variables in the database such as past donation behavior, past volunteer activity, candidate support score, predicted household wealth, and Census-based neighborhood characteristics (Tam Cho and Gimpel 2007). It can then develop a fundraising behavior score that predicts the expected return for a call to a particular citizen. These scores are probabilistic, and of course it would be impossible to only call citizens who would donate $60, but large gains can quickly be realized. For instance, if a fundraising score eliminated half of the calls to citizens who would donate nothing, so that in the resulting distribution would be 30 percent donate $0, 35 percent donate $10, 17.5 percent donate $20, and 17.5 percent donate $60. The expected revenue from each call would increase from $10 to $17.50. Fundraising scores that increase the proportion of big donor prospects relative to small donor prospects would further improve on these efficiency gains.

The same logic can be applied to target expenditures for voter mobilization and persuasive communications. Targeting persuasive communications to citizens who are extremely unlikely to vote is inefficient. Even if the persuasive communication were effective at convincing these citizens to support the campaign's candidate or issue, the usual assumption among practitioners is that changing citizens' candidate or issue preferences does not meaningfully change their likelihood of voting. A similar logic could be applied to citizens who are already extremely likely to support a campaign's candidate or issue. If the support score predicts that a citizen is 98 percent likely to support a campaign's candidate or issue, and assuming the opposing campaign's activities will not meaningfully undermine this citizen's support likelihood, one might decide that persuasive communications would be better targeted to citizens who have a moderate or low likelihood of supporting the campaign's candidate or issue, along with a high likelihood of voting. Relying on turnout scores and support scores to target persuasion efforts in this manner represents an increase in efficiency just as fundraising scores improve the cost effectiveness of fundraising calls.

The value of using predictive scores for targeting has become widely recognized by campaigns during the past five years. Sophisticated use of these predictive scores allows campaigns to simultaneously broaden the populations targeted while pruning away groups they believe will be cost ineffective.

Catalist, LLC, is a political data vendor that compiles and maintains nationwide registration, demographic, and other political data for progressive, civic, and non-profit organizations such as labor unions, political candidates, and other advocacy groups. They build predictive scores using this data to help their clients analyze the electorate and target their activities more efficiently. The firm provided data for showing how its targeting of populations for its clients evolved over the last three presidential elections in Ohio (see Ansolabehere and Hersh 2010). The discussion that follows includes data from the Kerry campaign in 2004 and the Obama campaign in 2008 and 2012 Ohio candidates other than Obama. In each election, Catalist had several hundred clients across the state of Ohio. Catalist categorizes potential Ohio voters along two scales: whether or not they are likely to vote, and whether they are more likely to vote Democratic, Republican, or in-between. Divide each of these measures into a scale with 50 gradations, making a total of 2500 different cells. You can then create a heat map of how often each one of those cells is contacted by allied campaigns, including all modes of contact for all purposes across the election cycle (see on-line appendix). Given the centrality of Ohio in the past three Presidential elections, the calculations represent tens of millions of voter contacts.

Although Catalist's client base differed across all three cycles, this analysis shows the increasing reliance on targeting scores for their collective voter targeting efforts. In 2004, when few clients relied on predictive scores for targeting, Catalist found that most contact was concentrated among people predicted to support Democratic candidates, regardless of their likelihoods of voting. This meant that campaign resources were probably inefficiently allocated, with a substantial share going to Democrats who were extremely unlikely to vote, or to Democrats who were extremely likely to vote and did not require either mobilization or persuasion. In 2008, Catalist clients appear to have relied more on predictive scores for their targeting. The highest concentrations of direct contacts were observed among citizens who were predicted to support Democratic candidates but who had low likelihoods of voting, i.e., those who might be reasonable targets for voter mobilization. They also targeted high turnout citizens with middling partisanship scores, who might be reasonable targets for "persuasion." The reasonableness of targeting in these ways depends on the likelihood that voters can be moved to turn out, or to be persuaded. As mentioned above, a current practice is to develop "responsiveness scores" based on pilot experiments to optimize targeting – particularly for persuasion outreach. As a result, the targeting in 2008 appears much closer to optimal than was observed in 2004. The results for 2012 look much the same as those of 2008 except with smoother transitions and more consistency across the landscape, suggesting even wider adoption of predictive scores for targeting. One noticeable difference between the 2012 figure and those of previous cycles is that Catalist clients appear to have avoided communicating with citizens with the lowest turnout probabilities. Catalist's clients may have chosen this strategy for a range of reasons, but regardless of their strategic reasons, apparently Catalist's 300-plus Ohio clients in 2012 used predictive scores to manifest their strategic plans in ways that they had not in previous cycles.

What Are Predictive Scores Worth?

Campaign organizations have adopted predictive scores, which suggests that they are electorally useful. They use these scores to target nearly every aspect of campaign outreach: door-to-door canvassing; direct mail; phone calls; email; television ad placement; social media outreach (e.g., Facebook and Twitter); and even web page display. Determining exactly how much using these scores affects electoral outcomes is difficult because the counterfactual is unclear. Is the appropriate comparison for assessing the value of campaign analytics between the current uses of predictive scores for targeting and a complete absence of targeting? Or would it be to compare currents uses to the basic heuristics that were used for targeting in the relatively recent past? Regardless, it is possible to derive bounds as to how much campaign analytics could matter to campaigns.

Persuasive communications is a good place to begin because targeting is so diffuse. There are so many possible targets, including potentially all citizens, and so many strategies, from shoring up support to causing opposition supporters to defect. Thus, persuasive campaign outreach can be directed almost anywhere along the support score spectrum from hard-core supporters to hard-core opponents. Thus, many campaigns use responsiveness scores as part of targeting their persuasive communications (Issenberg 2012a,b,c). Suppose a campaign's persuasive communications has an average treatment effect of 2 percentage points – a number on the high end of persuasion effects observed in high-expense campaigns: that is, if half of citizens who vote already planned to vote for the candidate, 52 percent would support the candidate after the persuasive communication. If a campaign indiscriminately attempted to persuade 8,500,000 citizens –about the size of the Florida electorate – it would generate 170,000 votes under this scenario.

Now imagine that the campaign has created a responsiveness score that predicts which citizens would be most responsive to its persuasive communications. Based on the responsiveness score, those in the top quintile are three times more responsive to the persuasive communications than the average citizen, the next quintile is twice as responsive, the middle quintile is no more responsive than average, the second quintile shows no average responsiveness to the persuasive communications, and the bottom quintile actually exhibited backlash to the persuasive communications equal to the overall average treatment effect. Table 1 illustrates these outcomes.[4] Actual campaign data analysts would construct a continuous responsiveness score, but this example involving quintiles suffices for illustration.

For campaigns with the resources to contact only 20 percent of the electorate, the responsiveness score allows them to create 102,000 votes (1,700,000 x 0.02 x 3 = 102,000). Without any form of targeting the campaign would generate only 34,000 votes (1,700,000 x 0.02 = 34,000), so using predictive scores doubles the number of new votes (see Table 1, row 1). A better financed campaign that could contact 40 percent of the electorate and would target the two most promising quintiles of the population. This strategy would yield a total of 170,000 votes, which is a 150 percent increase over having no targeting (3,400,000 x 0.02=68,000) (see Table 1, row 2). In this scenario, using predictive scores still improves the campaign's impact, but the gain is less than that of the more resource-constrained campaign. A campaign with the resources to push up against the zero bound where additional contacts begin to cost the campaign votes would see its efficiency improve by only 50 percent (see Table 1, row 4). This dynamic means that smaller campaigns will benefit most from targeting based on predictive scores, but they are the ones who are least able to afford hiring campaign data analysts and voter databases. Well-financed campaigns benefit from targeting based on predictive scores, but yield smaller relative gains over not using predictive scores for targeting. In this sense, given that small campaigns tend to be less reliant on data analytics, it appears that smaller campaigns are under investing in the development and use of predictive scores.

[4] Backlash is not an uncommon observation among field experiments examining persuasive campaign effects (for example, Arceneaux and Kolodny 2009; Bailey, Hopkins and Rogers 2013), and among other types of experiments (Nicholson 2012; Hersh and Shaffner 2013)

Table 1: Hypothetical Example of Persuasion Responsiveness Score's Value

Quintile	Effect Multiplier	Votes created in quintile	Cumulative votes	Improvement over no targeting
Top 20%	3	102,000	102,000	200%
60 – 80%	2	68,000	170,000	150%
Middle 20%	1	34,000	204,000	100%
20 – 40%	0	0	204,000	50%
Bottom 20%	-1	-34,000	170,000	20%

Note: This example assumed that the average effect of campaign contact is 2 percentage points, and that the electorate size is 8,500,000.

Again using the fairly generous multiplier regarding responsiveness scores and a baseline 2 percentage point average treatment effect, we can set an upper bound on how the use of such a score might affect campaign outcomes. If there are 8,500,000 citizens who will vote in a state (roughly the number of votes cast in the 2012 presidential election in Florida), and a campaign can successfully administer the attempted direct persuasive communications to only half the targeted citizens because of inability to reach all citizens, then a campaign that does not use responsiveness scores would generate 85,000 votes while a campaign that uses responsiveness scores would generate 102,000 votes through direct persuasive communications. While the difference of 17,000 votes is notable, it constitutes only 0.2 percent of the overall vote in this jurisdiction. That said, it would have constituted 23 percent of the 74,309 vote margin of victory for the Obama campaign in 2012.

Campaigns do not want to mobilize citizens to vote who support their opponent, so one of the most important uses for support scores is to identify which citizens should be targeted during voter mobilization efforts. In an evenly divided electorate, indiscriminately mobilizing citizens would net zero votes—because as many opponents would be mobilized as supporters. In this setting, a naïve comparison of data-based campaigning to absolutely no targeting is not appropriate. Instead, consider a comparison with the following relatively basic targeting strategy that is still employed today in electoral settings that do not have access to predictive scores. Imagine that a campaign attempts to identify individual citizens who support their candidate or issue by directly contacting them in person or over the phone. Imagine that this campaign can successfully reach half of the population and accurately identify their candidate/issue preference. For the remaining half of the population for whom the campaign has not identified a preference, the campaign proceeds to sweep through neighborhoods where more than half of the population supports the campaign's candidate, on the assumption that this approach will lead to a net gain in votes. The only people not targeted in these sweeps are those individuals concretely identified as supporters of the opponent. We can therefore express the expected yield in votes from this targeting strategy as

$$\begin{cases} 0.5\beta N_j(\%Support_j) \text{ if } \%Support_j < 0.5 \\ \beta N_j(\%Support) - 0.5\beta N_j(\%Oppose) \text{ if } \%Support_j > 0.5 \end{cases}$$

where β, is the mobilization effect from the campaign, $\%Support_j$ is the level of support for the candidate in precinct *j*, and N_j is the number of registered voters in precinct *j*.

The first line points out that in precincts where support for the candidate is less than 50 percent, the only effect of this plan will be the direct contacts with supportive voters. However, by assumption the campaign only has the ability to identify half of these people. The second line points out that in areas where support for the candidate is more than 50 percent, the strategy will have two effects. The first is the benefit from mobilizing supporters in the precinct. Unfortunately, the sweep also mobilizes opponents in the proportion to which they are present ($\%Oppose$). However, the campaign managed to identify half of the people supporting the opposition and can choose to avoid these individuals, so the counter-productive mobilization can be cut in half.

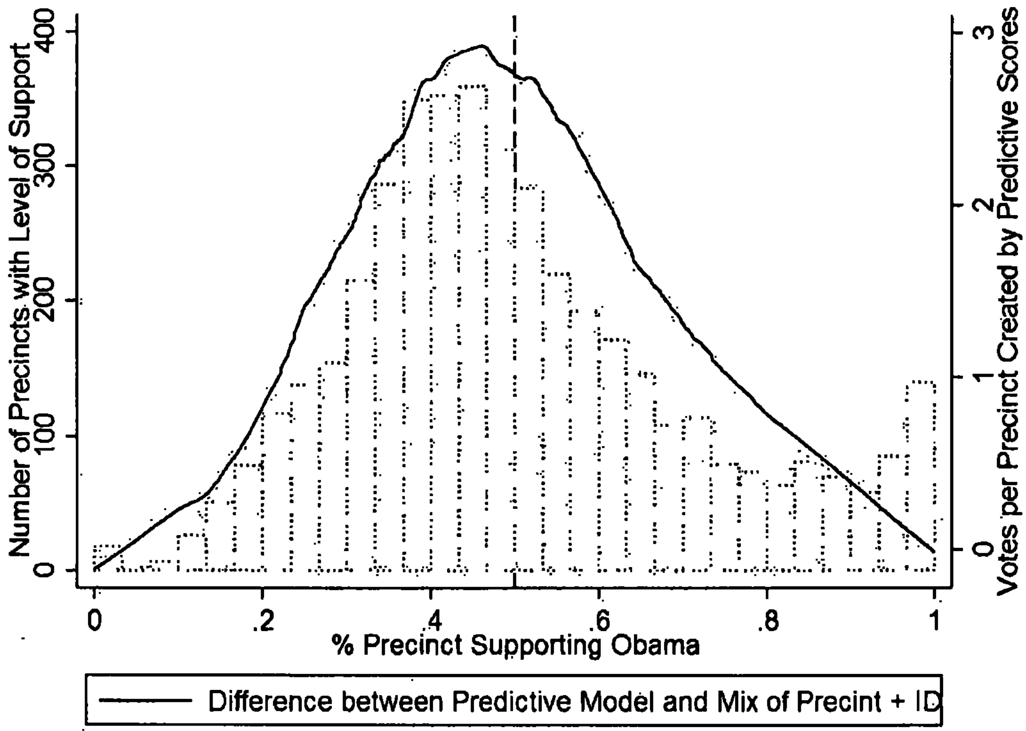
We can now contrast this targeting strategy to an imagined predicted support score strategy. It would obviously be an unfair comparison to argue that the predicted support score strategy worked without error, so we assume that it includes both false positives (misidentifying opponents as supporters) and false negatives (misidentifying supporters as opponents). One can think of these errors as reflecting the political diversity of a given neighborhood. In precincts where the vote is split 50/50, the false positive and false negative error rates are both 15 percent, because these would be the precincts where it is most difficult to infer political beliefs. However, in this hypothetical example the error rate tapers linearly as the precinct becomes more informative of resident beliefs, so that if a precinct unanimously supports one candidate or another, the error rate would obviously be zero. The equation below presents the formula used in this hypothetical model:

$$\begin{cases} \beta N_j[\%Support_j(0.85) - \%Oppose_j(0.15)] \text{ if } \%Support_j = 0.50 \\ \beta N_j[\%Support_j(1-0.15*\frac{\%Support_j}{0.5}) - \%Oppose_j * 0.15\frac{\%Oppose_j}{0.5}] \text{ if } \%Support_j < 0.50 \\ \beta N_j[\%Support_j(1-0.15*\left(1-\frac{\%Support_j}{0.5}\right) - \%Oppose_j * 0.15*\left(\frac{1-\%Oppose_j}{0.5}\right)] \text{ if } \%Support_j > 0.50 \end{cases}$$

The equations make clear one under-appreciated aspect of predictive modeling; modeling can only increase the efficiency of mobilization efforts. If the outreach from the campaign is not effective (i.e., $\beta = 0$), then no votes are generated. Big data analytics may receive media attention, but its effectiveness is entirely reliant on the strength of more traditional aspects of the campaign. If a campaign does not have effective outreach to voters, then predictive analytics cannot solve that problem.

Comparing the traditional strategy of "identification and sweep" to the predictive model, two advantages of the predictive model become clear. First, predictive analytics allows the campaign to target likely supporters in otherwise unfriendly territory. Before accurate prediction was possible, campaigns would leave votes on the table by ignoring supporters living in opponent strongholds. Given the expense of actually identifying individual voter's preferences and the relatively low yield of supporters, avoiding these areas was not optimal tactically, but understandable. Second, precinct sweeps are inefficient because in evenly divided precincts many non-supporters are also mobilized and thereby decrease the overall effectiveness of mobilization drives. Predictive scores (to the extent they are accurate) can prevent this inefficiency. As a result, conditional on precinct size, the biggest difference between the traditional "identification and sweep" tactic and modeled scores is found in the most evenly divided precincts.

Figure 2: Difference between Predictive Scores and Older campaign targeting heuristics.



Note: X-axis is percent of the two-party vote share for Obama in a precinct in the 2012 general election.
Left y-axis, represented by dotted bars, reports the number of precincts with that given level of support for Obama.
Right y-axis, represented by the solid line, reports the hypothesized difference between the use of predictive scores for targeting and the use of "identification and sweep." Beta is assumed to be 0.01.
The distribution of precinct data comes from all 4,354 precincts in the 2012 presidential election in Florida.

Figure 2 shows the results of a thought experiment if these two tactics had been used in Florida across all 4,354 precincts during the 2012 election. The x-axis depicts the percent of votes cast in favor of President Obama in each precinct and the left-hand y-axis shows in how many precincts President

Obama received that share of the vote. Thus, President Obama received between 0 and 3 percent of the vote in about 20 precincts (the left-most bar), and received between 97 and 100% of the vote in 140 precincts (the right-most bar). Now imagine as a hypothetical example that the Obama campaign knows the distribution of its support across precincts before the election, and is considering two possible strategies to increase its vote: the old-style "identification and sweep" combination of direct contact and precinct targeting, or the method using prediction scores. The solid line, measured on the right y-axis, shows the difference in the number of votes generated from these two approaches. The biggest difference between the two strategies takes place in the middle of the distribution where precincts are most evenly split.[5] The reason for this is clear when the tails are considered. In areas where support for Obama was low, there were not many Obama supporters to mobilize. In the areas where support for Obama was high, there were many supporters to mobilize, but both targeting strategies would target these citizens and neither would mistakenly mobilize those who support the opposing campaign's candidate. It is in areas where the precinct-level data is not predictive of which candidate citizen's support where predictive scores at the individual-level—even given the built-in assumption of a higher number of false positives and false negatives in these precincts--yield the greatest value.

With these assumptions, we can gain a rough sense of the impact of the Obama 2012 mobilization effort in Florida using the predictive scores for targeting (which was the strategy the campaign reportedly employed) compared to a precinct-based targeting strategy. Assuming the campaign had a 1 percentage point effect on turnout among the half of the citizens that it targeted for mobilization and successfully contacted, we estimate that it would have generated 8525 more votes in Florida targeting based on predictive scores relative to targeting based on precinct. This vote total would have been decisive in the 2000 election between Bush and Gore, and still constitutes 11 percent

[5] If the number of registered voters was held constant across precincts, then the point of maximum difference would be at 0.5. However, the precincts where Obama received 42 – 45% of the vote are larger than precincts with an even split so there are more votes to be harvested just to the left of the 50/50 mark.

of the 74,309 vote margin of victory Obama enjoyed in 2012. Combined with the persuasion analysis above, this thumbnail sketch makes an argument that the 2012 would have been closer in key states had it used the older and coarser targeting technologies, rather than the predictive scores produced by its campaign data analysts.

Conclusion: Some Thoughts on Coordination

Sophisticated campaigns develop and use voter databases that contain a range of detailed information on individual citizens. As a result, campaign data analysts occupy an increasingly important role in politics. They develop predictive models that produce individual-level scores that predict citizens' likelihoods of performing certain political behaviors, supporting candidates and issues, and responding to targeted interventions. The use of these scores has increased dramatically during the last few election cycles. Simulations suggest that these advances could yield sizable and electorally meaningful gains to campaigns that harness them.

Since predictive scores make campaigns more effective and efficient by increasing the cost effectiveness of communicating with citizens, a broad range of organizations do and will employ the technologies. To the extent that predictive scores are useful and reveal true unobserved characteristics about citizens, it means that multiple organizations will produce predictive score that recommend targeting the same sets of citizens. For example, some citizens might find themselves contacted many times while other citizens—like those with low turnout scores in 2012--might be ignored by nearly every campaign. The marginal effect of the fifth or sixth contact from a campaign will be less than the marginal effect of the first contact from a campaign. Thus, concentrating attention on the same set of citizens due to widespread adoption of predictive scores may offset some of the gains reaped from developing predictive scores in the first place. In this way, developing and using predictive scores creates a coordination game in which allied organizations would prefer to partition the electorate and not to duplicate efforts.

Coordination could theoretically happen between partisan organizations, like state parties, candidate campaigns, and coordinated campaigns, and across non-partisan activities, like civil rights groups, labor unions, and environmental groups. However, partisan and non-partisan organizations are not allowed to coordinate their electoral activities. Since it is nearly impossible to observe whom campaigns target for direct communications—that is, direct mail, knock on doors, and making phone calls—this coordination game has incomplete information, which means that inefficiencies from overlapping contacts are inevitable.

Even when coordination is allowed by law, coalitions may have conflicting incentives. There is enough regional variation in ideology that it is possible for local candidates to appeal to citizens who oppose the national candidate. For instance, local Republicans mobilizing citizens in liberal districts would have hurt Mitt Romney and local Democrats mobilizing citizens in conservative districts would have hurt Obama in 2012. The same dynamic plays out among non-partisan groups as well. While labor union members and environmentalists agree on many policies and values, it is likely that some members do not hold that same views on both labor and environmental issues. In states like West Virginia where the local industry (i.e., coal) is considered "dirty" by environmentalists, the groups could be working cross-purposes both with regards to messaging and targeting. Thus, mobilizing a set of citizens for a labor related ballot initiative might result in less support for an environmentally friendly candidate. This tension is endemic to the very nature of the federal system of representation and coalition politics. The tension has always been present, but now that groups can share very detailed targeting plans and support scores, the tension can and will bubbles to the surface more often than in the past.

The improved capability to target individual voters offers campaigns an opportunity to concentrate their resources where they will be most effective. This power, however, has not radically transformed the nature of campaign work. One could argue that the growing impact of data analytics in campaigns has amplified the importance of traditional campaign work. Message polling no longer solely dictates targeting, but the increased demand for information on campaign has increased the amount of polling used to generate snapshots of the electorate. Professional phone interviews are still used for message development and tracking, but they are also essential for developing predictive scores of candidate support and measuring changes in voter preferences in experiments. Similarly, better targeting has made grassroots campaign tactics more efficient and therefore more cost competitive with mass communication forms of outreach. Volunteers still need to persuade skeptical neighbors, but they are now better able to focus on persuadable neighbors and use messages more likely resonate. This leads to higher quality interactions and (potentially) a more pleasant volunteer experience. So while savvy campaigns will harness the power of predictive scores, the scores will only help the campaigns that were already effective.

References

Ansolabehere, Stephen, and Eitan Hersh. 2010. "The quality of voter registration records: A state-by-state analysis." *Cambridge, Mass.: Department of Government, Harvard University.*

Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20(4):437-459.

Arceneaux, Kevin, and Robin Kolodny. 2009. "Educating the least informed: Group endorsements in a grassroots campaign." *American Journal of Political Science* 53.4: 755-770.

Arceneaux, Kevin and David W. Nickerson. 2009. "Who is Mobilized to Vote? A Re-Analysis of Seven Randomized Field Experiments." *The American Journal of Political Science* 53(1):1-16.

Arceneaux, Kevin and David W. Nickerson. 2010. "Negative and Positive Campaign Messages: Evidence from Two Field Experiments." American Politics Research 38(1): 54-83.

Bailey, Michael A., Daniel J. Hopkins and Todd Rogers. 2013. "Unresponsive and Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts." *Working Paper.* http://ssrn.com/abstract=2307631 or http://dx.doi.org/10.2139/ssrn.2307631

BigData-Startups. 2013. "Big Data during the Obama campaign." http://www.bigdata-startups.com/BigData-startup/big-data-obama-campaign/. Accessed February 18, 2014.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen. 1984. *Classification and Regression Trees.* Boca Raton, FL: CRC Press.

Breiman, Leo. 2001. "Random Forests". *Machine Learning* 45:5–32.

Bryan, Christopher J., Gregory M. Walton, Todd Rogers, and Carol S. Dweck. 2011. "Motivating voter turnout by invoking the self." *Proceedings of the National Academy of Sciences 108*(31): 12653-12656.

Burt, Gabriel. 2013. "My Time at 'Obama for America'." *Tech in Motion* August 27, 2013. http://www.techinmotionevents.com/blog/post/2013/08/27/my-time-at-obama-for-america.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1):266–98.

Dietterich, Tom. 1995. "Overfitting and Undercomputing in Machine Learning." *ACM Computing Surveys* 27(3):326-327.

Duhigg, Charles. 2012. "Campaigns Mine Personal Lives to Get Out Vote." *New York Times,* October 14, 2012, page A1.

Gallagher, Sean. 2012. "Built to win: Deep inside Obama's campaign tech" Arstechnica.com, November 14, 2012. http://arstechnica.com/information-technology/2012/11/built-to-win-deep-inside-obamas-campaign-tech/

Gan, Guojun, Chaoqun Ma, and Jianhong Wu. 2007. *Data Clustering: Theory, Algorithms, and Applications.* Alexandria, VA: SIAM, Society for Industrial and Applied Mathematics.

Gerber, Alan, and Todd Rogers. 2009. "Descriptive social norms and motivation to vote: Everybody's voting and so should you." *Journal of Politics* 71: 1-14.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3):653-63.

Gerber, Alan S. and Donald P. Green. 2008. *Get Out the Vote!* Washington, DC: Brookings Institution Press.

Green, Donald P. and Holger L. Kern. 2011. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491-511.

Haberman, Maggie. 2013. "Obama '12 vets to share some, not all, data." *Politico* November 20, 2013. http://www.politico.com/story/2013/11/barack-obama-2012-campaign-data-100133.html

Hersh, Eitan D. and Brian F. Schaffner. 2013. "Targeted Campaign Appeals and the Value of Ambiguity." *Journal of Politics 75 (2): 520-534.*.

Hillygus, D. Sunshine and Todd G. Shields. 2008. *The Persuadable Voter: Wedge Issues in Presidential Campaigns*. Princeton, NJ: Princeton University Press.

Issenberg, Sasha. 2010. "Nudge the Vote: How Behavioral Science is Remaking Politics." *New York Times Sunday Magazine*, October 31, 2010, page MM28.

Issenberg, Sasha. 2012a. *The Victory Lab: The Secret Science of Winning Campaigns*. New York, NY: Crown Publishers.

Issenberg, Sasha. 2012b. "The Death of the Hunch." *Slate*, March 22, 2012. http://www.slate.com/articles/news_and_politics/victory_lab/2012/05/obama_campaign_ads_how_the_analyst_institute_is_helping_him_hone_his_message_.html

Issenberg, Sasha. 2012c. "How President Obama's campaign used big data to rally individual voters, Part 2." *MIT Technology Review*, December 17, 2012. http://www.technologyreview.com/featuredstory/508851/how-obama-wrangled-data-to-win-his-second-term/

Kuhn, David Park. 2007. "Vacation Politics." CBSNews.com, December 5, 2007. http://www.cbsnews.com/2100-250_162-607234.html.

Madrigal, Alexis C. 2012. "When the Nerds Go Marching In." TheAtlantic.com, November 16, 2012. http://www.theatlantic.com/technology/archive/2012/11/when-the-nerds-go-marching-in/265325

Malchow, Hal. 2003. *The new political targeting*. Campaigns and Elections Magazine.

McAuliffe, Terry and Steve Kettman. 2008. *What A Party!: My Life Among Democrats: Presidents, Candidates, Donors, Activists, Alligators, and Other Wild Animals*. New York, NY: St. Martin's Press.

Nicholson, Steven P. 2012. "Polarizing cues." *American journal of political science 56*(1): 52-66.

Nickerson, David and Todd Rogers. 2010. "Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making." *Psychological Science* 21(2): 194-199.

Nickerson, David W. 2005. "Partisan Mobilization Using Volunteer Phone Banks and Door Hangers," The Annals of the American Academy of Political and Social Science 601:10-27.

Nickerson, David W. 2006a. "Hunting the Elusive Young Voter," *Journal of Political Marketing* 5(3):47-69.

Nickerson, David W. 2006b. "Volunteer Phone Calls Can Increase Turnout." *American Politics Research* 34(3):271-292.

Nickerson, David W. 2007a. "Quality is Job One: Volunteer and Professional Phone Calls," *American Journal of Political Science* 51(2):269-282.

Nickerson, David W. 2007b. "Don't Talk to Strangers: Experimental Evidence of the Need for Targeting." Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 12-15, 2007.

Nickerson, David W., Ryan F. Friedrichs, and David C. King. 2006. "Partisan Mobilization Experiments in the Field: Results from a Statewide turnout experiment in Michigan," Political Research Quarterly 34(1):271-292.

Rogers, Todd and Masa Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, And Is (Misleadingly) Unbiased." *American Politics Research*. doi: 10.1177/1532673X13496453

Rogers, Todd and David Nickerson. 2013. "Can Inaccurate Beliefs About Incumbents be Changed? And Can Reframing Change Votes?" *HKS Faculty Research Working Paper Series*, RWP13-018.

Scherer, Michael. 2012. "Inside the Secret World of Quants and Data Crunchers Who Helped Obama Win." *Time*, November 19, 2012, p. 56-60.

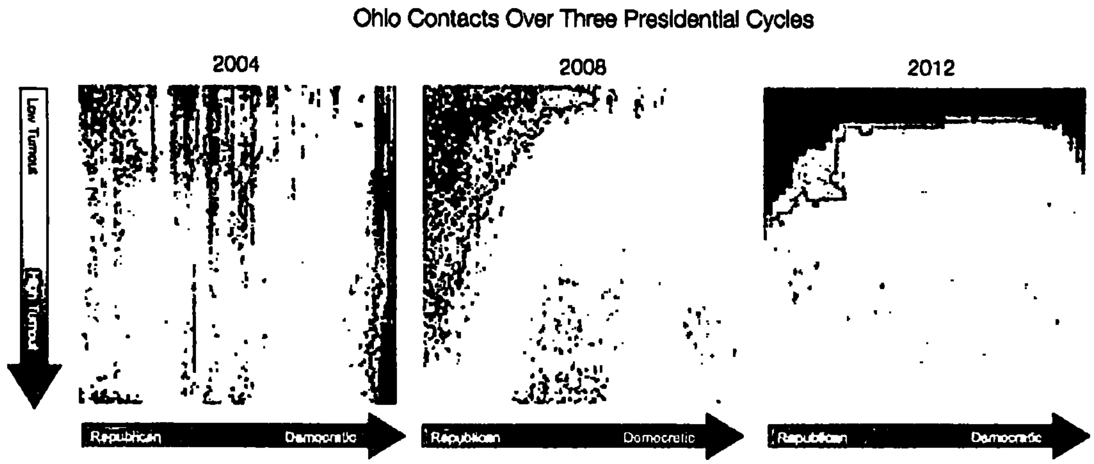
Tam Cho, Wendy K. and James G. Gimpel. 2007. "Prospecting for (Campaign) Gold." *American Journal of Political Science* 51(2):255-268.

Vargas, Jose Antonio. 2013. "Obama Raised Half a Billion Online." *Washingtonpost.com* "The Clickocracy" November 20, 2008. http://voices.washingtonpost.com/44/2008/11/20/obama_raised_half_a_billion_on.html

Appendix: An Illustration of Using Predictive Scores

As discussed in the text, this figure shows the pattern of contact with potential voters for Catalist political clients in Ohio during the 2004, 2008, and 2012 election cycles. Each panel shows the same heat map. The y-axis describes citizens' turnout behavior scores; citizens at the bottom are the most likely to vote in that election. The x-axis describes citizens' likelihoods of supporting Democratic candidates as opposed to Republican candidates; citizens who are likely to support Republican candidates are on the left and citizens who are likely to support Democratic candidates are on the right. Each axis is broken into 50 equally sized bins (2500 bins in total), and each bin is colored by the intensity of direct contact the average citizen in the bin received over the course of the election. This includes all modes of direct contact, for all purposes, across the entire election cycle. Darker green boxes were contacted at a relatively high rate; darker red boxes at a relatively lower rate; and shades of orange and yellow are in between.

Figure 1



Source: Catalist, LLC
X-axis is likelihood of supporting a Democratic candidate over a Republican candidate, ranging from 0 (left) to 100 (right).
Y-axis is likelihood of voting ranging, ranging from 100 (low) to 0 (high).
Colors represent density/frequency of direct contacts from all Catalist clients over the course of the entire election cycle. Dark red means these citizens received the fewest direct contacts over the election cycle, and dark green means these citizens received the most direct contacts over the election cycle.